Clifford Lam and Pedro C. L. Souza

Estimation and selection of spatial weight matrix in a spatial lag model

Article (Accepted version) (Refereed)

Estimation and Selection of Spatial Weight Matrix in a Spatial Lag Model

Clifford Lam*[1] and Pedro CL Souza†[2]

[1]Department of Statistics, London School of Economics and Political Science
[2]Department of Economics, University of Warwick

Abstract

Spatial econometric models allow for interactions among variables through the specification of a spatial weight matrix. Practitioners often face the risk of misspecification of such a matrix. In many problems a number of potential specifications exist, such as geographic distances, or various economic quantities among variables. We propose estimating the best linear combination of these specifications, added with a potentially sparse adjustment matrix. The coefficients in the linear combination, together with the sparse adjustment matrix, are subjected to variable selection through the adaptive Least Absolute Shrinkage and Selection Operator (LASSO). As a special case, if no spatial weight matrices are specified, the sparse adjustment matrix becomes a sparse spatial weight matrix estimator of our model. Our method can therefore be seen as a unified framework for the estimation and selection of a spatial weight matrix. The rate of convergence of all proposed estimators are determined when the number of time series variables can grow faster than the number of time points for data, while Oracle properties for all penalized estimators are presented. Simulations and an application to stocks data confirms the good performance of our procedure.

Key words and phrases. Spatial weight matrix; adaptive LASSO; spatial fixed effects; sparse adjustment; partial sign consistency; oracle property.

*Clifford Lam is Associate Professor, Department of Statistics, London School of Economics. Email: C.Lam2@lse.ac.uk

†Pedro CL Souza is Assistant Professor, Department of Economics, University of Warwick. Email: pedro.souza@warwick.ac.uk

1 Introduction

Spatial econometrics is the study of the interaction between units or entities. Such interactions come in many forms. Examples include the peer effects of learning in classroom (Ammermuller and Pischke, 2009, Angrist and Lang, 2004), the spread of crime and delinquent behavior (Glaeser et al., 1996) in criminology, and contagion between financial markets (Longstaff, 2010). In spatial econometric models, such interactions are usually assumed to be known through the so-called "spatial weight matrix", which is a square matrix of size N, with N being the number of time series variables that are being modeled.

It is inevitable to specify a spatial weight matrix when using a spatial econometric model. Yet, in some scenarios there is no data to do so, for example when social interactions between the units or entities are not available to the researcher. Even with enough data to do so, if N is too large relative to the sample size T, the parameters in the spatial weight matrix are not identified without further assumptions on the structure of the matrix itself (e.g., sparsity), or other forms of regularization in place. For instance, Bhattacharjee and Jensen-Butler (2013) propose to estimate the spatial weight matrix under the symmetric constraint, although they assume N to be finite. Pinkse et al. (2002) and Sun (2016) consider estimating a nonparametric function of some underlying "distance" variables for the spatial weight matrix parameters, which inherently assumes some form of smoothness in the function itself. This assumption of smoothness can be inaccurate, depending on the "distance" variable involved. For example, Germany and Japan can be close economic competitors with large geographical distance between them. The papers suggest including several distance measures, but only provide theories for including one. It is in fact difficult to estimate nonparametric smooth functions of higher dimension without specific assumptions on the function itself, on top of the smoothness assumption. Beenstock and Felsenstein (2012) consider using the sample covariance of the data to infer the values of the spatial weight matrix. However, sample covariance matrix can suffer from serious bias in the extreme eigenvalues when N is of the same order as T (see topics in random matrix theory from, e.g., chapter 5 of Bai and Silverstein (2009), who describe mathematically the bias in the largest and the smallest eigenvalues of the sample covariance matrix). Ahrens and Bhattacharjee (2015) and Lam and Souza (2015) assume sparsity in the spatial weight matrix, and, in turn, use penalization methods such as the Least Absolute Shrinkage and Selection Operator (LASSO) for a penalized estimator. Indeed, the spatial weight matrix can be sparse. In many applications, however, it may not be sparse enough for a satisfactory penalized estimator.

Aside from estimating the spatial weight matrix, another important approach in spatial econometrics is selecting an appropriate model through testing. Bailey et al. (2016) use multiple testing to infer if an element of the spatial weight matrix is zero, negative or positive. Liu and Prucha (2017) generalize the Moran $\mathcal{I}$ test to check if a linear combination of specified spatial weight matrices is appropriate for modelling the data under a particular spatial lag model, but they do not

consider the alternative hypothesis of an acceptable linear combination. Kelejian and Piras (2011) and Kelejian and Piras (2016) propose J-tests to find from among a number of alternatives with differently-specified spatial weight matrices. Clearly, the quality of the alternative hypothesis is very important, and pinpointing which spatial weight matrix to use may require multiple J-tests.

This paper tackles the the problem of spatial weight matrix estimation when multiple specifications of the spatial weight matrix are available, as Liu and Prucha (2017) and Kelejian and Piras (2016) consider. This is a very practical scenario. For instance, for geographical distance, one can specify w_{ij} in the spatial weight matrix $\mathbf{W}$ by calculating the inverse of the distance r^{-1} between units i and j. One can, in fact, use $r^{-\ell}$ with ℓ being a positive number to specify a spatial weight matrix. A simple adjacency matrix can also play the role of a spatial weight matrix. To combine information from these specifications, and select those that are useful to construct a spatial weight matrix, we consider finding their "best" sparse linear combination using the adaptive LASSO. This technique was proposed for variable selection in a linear regression model by Zou (2006). As a major contribution, we add a sparse adjustment matrix to the estimated best sparse linear combination, which overcomes the problem of potential misspecification of the spatial weight matrix. This sparse adjustment matrix, to be estimated from data, effectively incorporates errors of misspecification. Hence, if there are reasonable specifications, the sparse adjustment matrix is indeed expected to be truly sparse. In this sense, the sparsity assumption for the adjustment matrix is more easily satisfied than those in Ahrens and Bhattacharjee (2015) or Lam and Souza (2015), which require the spatial weight matrix itself to be sparse. If this sparse adjustment is estimated to be the zero matrix, it means that the specified spatial weight matrices are good fit for the data, so testing like that in Liu and Prucha (2017), or the J-test of Kelejian and Piras (2016), become unnecessary.

When no specified spatial matrices are available, the estimator of the sparse adjustment matrix itself becomes a sparse spatial weight matrix estimator. This is an important special case, as practitioners are allowed not to specify any spatial weight matrices. In this sense, our method provides a unified framework in spatial weight matrix selection and estimation. It allows us to estimate which specified spatial weight matrices are the most informative. Ultimately, it provides an estimator of the spatial weight matrix at the same time.

We assume a time-invariant and exogenous spatial weight matrix in this paper. Arnold et al. (2011) make the same assumptions and estimate a linear combination of three specified spatial weight matrices to model the stock returns of the Euro Stoxx 50 members. They estimate the coefficients using generalized method of moments, and assume a diagonal covariance matrix for the model errors with normality and serial independence. They also have a simple model without any regressors, but obtain good numerical results. In this paper, we generalize the model of Arnold et al. (2011) to include both exogenous or endogenous regressors and fixed effects, while allowing for non-Gaussian errors with more general serial dependence and covariance structure. Medeiros and Mendes (2016) explore the properties of LASSO and adaptive LASSO also under

non-Gaussian and conditionally heteroskedastic errors, but do not consider spatial dependence. We use exogenous covariates as instruments to uncouple intrinsic endogeneity of the observed variables similar to Fan and Liao (2014), although they use the Feasible Generalized Method of Moments (FGMM) method coupled with quasi-likelihood for the data.

We prove that the zero's in the sparse adjustment matrix and those in the coefficients of the spatial weight matrix linear combination can be selected consistently. We also prove the asymptotic normality for the non-zeros, as well as for the regression coefficient estimators in the spatial lag model. We can therefore carry out inference not only on the regression parameters, but also on the non-zero entries in the sparse adjustment matrix. This means that we can test how far individual spatial interactions are away from the "best" linear combination of the specified spatial weight matrices. We can also test how relevant each spatial weight matrix specification is, by looking at the estimated coefficients for the linear combination. In Section 5.2, we demonstrate that including more "relevant" spatial weight matrix specifications improves the inference and precision of the estimates of important regressors in the model. Incidentally, it shows that a better specification of spatial weight matrix helps us estimate the model parameters.

The rest of the paper is organized as follows. Section 2 presents the spatial lag model and the sparse adjustment idea on a linear combination of spatial weight matrices. The LASSO and adaptive LASSO estimation problems for various parameters in the model are also presented. Section 3 presents all the assumptions in the paper. We show sign consistency and asymptotic normality of our estimators, and identification of parameters asymptotically. Section 4 presents the algorithm used for calculating our estimators, and the BIC criterion is used to find suitable tuning parameters. Section 5 provides detailed simulations to demonstrate finite sample performance of our estimators. In particular, Section 5.2 provides a thorough empirical study on the cross-sectional dependence of stocks traded in the New York Stock Exchange (NYSE). Proofs of all theorems are given in the supplementary materials for this paper.

2 Model and Motivation

As a starting point, we consider the following spatial lag model with fixed effects,

$$\mathbf{y}_t = \boldsymbol{\mu}^* + \mathbf{W}^*\mathbf{y}_t + \mathbf{X}_t\boldsymbol{\beta}^* + \boldsymbol{\epsilon}_t, \quad t = 1, \ldots, T, \tag{2.1}$$

where $\mathbf{y}_t = (y_{1t}, \ldots, y_{Nt})^{\mathrm{T}}$ is an $N \times 1$ vector of dependent variables. The matrix $\mathbf{W}^*$ is an $N \times N$ spatial weight matrix with 0 on the main diagonal. It may have negative off-diagonal elements, and it is not necessarily symmetric. These features allow for both positive and negative, and possibly asymmetric, spatial interactions among the component time series. The vector $\boldsymbol{\mu}^*$ is an $N \times 1$ vector of constants (fixed effects), while the spatial regression parameter $\boldsymbol{\beta}^*$ has size $K \times 1$, so that the matrix of covariates $\mathbf{X}_t$ has size $N \times K$. The innovation process $\boldsymbol{\epsilon}_t$ has mean $\mathbf{0}$ and

covariance matrix $\mathbf{\Sigma}_\epsilon$. For more detailed assumptions, see Section 3.3.

In spatial econometrics, the matrix $\mathbf{W}^*$ is usually assumed to be known (up to an unknown multiplicative constant ρ^*, called the spatial autoregressive parameter). In this paper, we make use of possible expert knowledge regarding $\mathbf{W}^*$, while allowing the final estimator to deviate slightly from it. More specifically, we decompose $\mathbf{W}^* = \mathbf{A}^* + \rho^*\mathbf{W}_0$, where $\mathbf{W}_0$ is a pre-specified spatial weight matrix. We also denote $\mathbf{W}_0$ as an "expert spatial matrix" since it incorporates expert knowledge about the interactions between units, such as inverse distances between countries. The spatial autoregressive parameter ρ^* adjusts the magnitude and direction of the spatial interactions specified in $\mathbf{W}_0$. As in the spatial econometrics literature, we assume that $|\rho^*| < 1$ to ensure that the model is stationary. We also assume that the diagonal elements of $\mathbf{A}^*$ are 0.

If $\mathbf{W}_0$ exactly represents the true underlying spatial interaction pattern, then $\mathbf{A}^* = \mathbf{0}$. If $\mathbf{W}_0$ is close to the true underlying spatial interaction pattern, then $\mathbf{A}^*$ may not be exactly $\mathbf{0}$, but is expected to be sparse or approximately sparse (i.e., with many "small" elements). Hence, $\mathbf{A}^*$ can be interpreted as a sparse adjustment to the matrix $\mathbf{W}_0$, and model (2.1) becomes

$$\mathbf{y}_t = \boldsymbol{\mu}^* + (\mathbf{A}^* + \rho^*\mathbf{W}_0)\mathbf{y}_t + \mathbf{X}_t\boldsymbol{\beta}^* + \boldsymbol{\epsilon}_t, \quad t = 1, \ldots, T. \tag{2.2}$$

Finally, the special case $\mathbf{W}_0 = \mathbf{0}$ corresponds to sparse spatial weight matrix estimation without any specifications. The matrix $\mathbf{A}^*$ now plays solely the role of the spatial weight matrix for model (2.2). As outlined in the introduction, some papers attempted to estimate the spatial weight matrix directly from data, completely avoiding its specification. In this paper, we take pure spatial weight matrix estimation as a special case. We consider that elements of $\mathbf{A}^*$ are constants, both invariant over time and exogenously determined, so they are independent of $\mathbf{y}_t$ or $\boldsymbol{\epsilon}_t$. For example, $\mathbf{A}^*$ could be the fixed measure of proximity between pairs of countries that have not been captured by the pre-specified $\mathbf{W}_0$.[1]

As a further generalization, which can be of practical use to applied researchers, we assume that more than one potential spatial weight matrix can be specified. This is particularly relevant if the researcher has various options for specifying $\mathbf{W}_0$. With sparse adjustment in mind, we propose to decompose the true spatial weight matrix $\mathbf{W}^*$ into the following:

$$\mathbf{W}^* = \mathbf{A}^* + \sum_{i=1}^{M} \delta_i^* \mathbf{W}_{0i}, \quad -1 \leq \rho^* = \sum_{i=1}^{M} \delta_i^* \leq 1. \tag{2.3}$$

Here $\mathbf{A}^*$ is the sparse adjustment described before. The spatial autoregressive parameter ρ^* is now the sum of δ_i^*, $i = 1, \ldots, M$, when $\mathbf{A}^* = \mathbf{0}$. It can be considered a generalization of the traditional definition of spatial autoregressive parameter if $\mathbf{A}^* \neq \mathbf{0}$. The specification $\mathbf{W}_0$ described before is now replaced by a linear combination of $\mathbf{W}_{0i}$, $i = 1, \ldots, M$, so that M denotes the total number of spatial matrices. This way, we can perform penalized estimation for the δ_i^*'s (see model (2.7)

[1] Only recently has the literature considered dynamic neighboring matrices (see Elhorst (2014), among others).

below) to see which specification, or linear combination of them, is the best suited to the data.[2] From the real data analysis in Section 5.2, including important specifications in the model can significantly improve the estimation of the spatial regression coefficients, and our penalization procedure facilitates this.

With sparse adjustment to a linear combination of specified spatial weight matrices, the complete model then reads

$$\mathbf{y}_t = \boldsymbol{\mu}^* + \left(\mathbf{A}^* + \sum_{i=1}^{M} \delta_i^* \mathbf{W}_{0i}\right)\mathbf{y}_t + \mathbf{X}_t\boldsymbol{\beta}^* + \boldsymbol{\epsilon}_t, \quad t = 1, \ldots, T. \tag{2.4}$$

Model (2.4) is a generalization of the usual spatial lag model formulation. See model (7.22) in LeSage and Pace (2008), for example. For the rest of the paper, we focus on analyzing model (2.4), since model (2.2) is, in fact, a special case of (2.4). In what follows, we assume without loss of generality that $E(\mathbf{X}_t) = \mathbf{0}$. Otherwise, we can write

$$\mathbf{X}_t\boldsymbol{\beta}^* + \boldsymbol{\mu}^* = (\mathbf{X}_t - E(\mathbf{X}_t))\boldsymbol{\beta}^* + (\boldsymbol{\mu}^* + E(\mathbf{X}_t)\boldsymbol{\beta}^*),$$

so that the spatial fixed effects are now captured by $\boldsymbol{\mu}^* + E(\mathbf{X}_t)\boldsymbol{\beta}^*$ rather than $\boldsymbol{\mu}^*$, and the covariates have mean $\mathbf{0}$.

2.1 Instruments and the augmented model

Since $\mathbf{y}_t$ (and possibly $\mathbf{X}_t$) is correlated with $\boldsymbol{\epsilon}_t$, we propose an instrumental variable version of model (2.4). We assume that instruments $\mathbf{U}_t$ for $t = 1, \ldots, T$, each of size $N \times \ell$ with $\ell \geq 1$, are available to the researcher. Each $\mathbf{U}_t$ is independent of $\boldsymbol{\epsilon}_t$, but is correlated with $\mathbf{y}_t$ in general (and $\mathbf{X}_t$, if $\mathbf{X}_t$ is also endogenous). Hence, $\mathbf{U}_t$ serves as instruments for model (2.4), since it only correlates with $\mathbf{y}_t$ and $\mathbf{X}_t$ but not $\boldsymbol{\epsilon}_t$. If $\mathbf{X}_t$ is exogenous, then $\mathbf{U}_t = \mathbf{X}_t$.

Following Kelejian and Prucha (1998), we generate instruments from within the model by interacting $\mathbf{U}_t$ with the expert $\mathbf{W}_{0i}$, $i = 1, \ldots, M$. Define the matrix $\mathcal{B}_t$ as

$$\mathcal{B}_t \;=\; \{\mathbf{U}_t, \mathbf{W}_{01}\mathbf{U}_t, \mathbf{W}_{01}^2\mathbf{U}_t, \ldots, \mathbf{W}_{0M}\mathbf{U}_t, \mathbf{W}_{0M}^2\mathbf{U}_t, \ldots\}, \tag{2.5}$$

which contains valid instruments since spatial lags of $\mathbf{U}_t$ are independent of $\boldsymbol{\epsilon}_t$. We denote $\mathbf{B}_t$ as a matrix of linearly independent vectors from (2.5), with dimension $T \times L$, where $L \geq K$.[3] Note that if $M = 0$, meaning that no expert spatial weight matrices are specified, it is not possible to

[2] One possible extension is to include several weight matrices to accommodate a structural break if breakpoint t^* is known. For instance, the researcher could specify two separate $1[t \leq t^*]\mathbf{W}_{0i}$ and $1[t > t^*]\mathbf{W}_{0i}$. This relaxes the assumption that $\mathbf{W}^*$ is time-invariant in a limited but potentially revealing way. We leave this extension for future work.

[3] Ideally, we should have instruments of the form $(\mathbf{A}^* + \mathbf{W}^*)^k\mathbf{U}_t$ for $k = 0, 1, 2, \ldots$. However, we do not know $\mathbf{A}^*$ and δ_i^*, and hence we select terms for which the values are known, excluding any cross-terms involving more than one $\mathbf{W}_{0i}$.

calculate the spatial lags of $\mathbf{U}_t$, and then $\mathbf{U}_t$ must have size $N \times L$ with $L \geq K$.

To utilize $\mathbf{B}_t$, we need to find a linear combination of instruments as correlated to the endogenous variables as possible. However, in what follows we compute the simple average of the vectors in $\mathbf{B}_t - \bar{\mathbf{B}}$, where $\bar{\mathbf{B}} = T^{-1}\sum_{t=1}^{T}\mathbf{B}_t$. The final vector of instruments is then $(\mathbf{B}_t - \bar{\mathbf{B}})\boldsymbol{\gamma}$, where $\boldsymbol{\gamma} = L^{-1}\mathbf{1}_L$ and $\mathbf{1}_L$ is the $L \times 1$ vector of ones. The weights vector $\boldsymbol{\gamma} = L^{-1}\mathbf{1}_L$ is a simple way to aggregate the instruments. This is done for clarity of exposition only, and can be estimated from the data as well. We defer the discussion about the optimal choice of instruments to Section 4.1. The proofs of all our theorems are identical irrespective of the method for choosing $\boldsymbol{\gamma}$.

2.2 Penalized estimators - LASSO

We perform penalized estimation of $\mathbf{A}^*$ since it is assumed to be sparse. This can be achieved regardless of whether $\mathbf{W}_{0i}$ are good specifications. If $\mathbf{W}_{0i} = \mathbf{0}$ for all i then we are estimating a sparse spatial weight matrix $\mathbf{A}^*$. If $\mathbf{A}^*$ is approximately sparse, "small" estimated elements of $\mathbf{A}^*$ are shrunk to 0. At the same time, we want to select which of the $\mathbf{W}_{0i}$'s contribute to the spatial weight matrix in (2.3). Again, a penalized estimation of δ_i^* serves exactly this purpose.

We start by profiling out $\boldsymbol{\beta}$. If $\mathbf{A}$ and δ_r are given, model (2.4) becomes

$$(\mathbf{I}_N - \mathbf{A} - \sum_{r=1}^{M}\delta_r\mathbf{W}_{0r})\mathbf{y}_t = \boldsymbol{\mu}^* + \mathbf{X}_t^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{\epsilon}_t, \quad t = 1, \ldots, T.$$

where the true values $\mathbf{A}^*$, δ_r^* and $\boldsymbol{\beta}^*$ are replaced by $\mathbf{A}$, δ_r and $\boldsymbol{\beta}$ respectively. Multiplying both sides by $(\mathbf{B}_t - \bar{\mathbf{B}})^{\mathrm{T}}$ and summing over t, we obtain

$$\sum_{t=1}^{T}(\mathbf{B}_t - \bar{\mathbf{B}})^{\mathrm{T}}(\mathbf{I}_N - \mathbf{A} - \sum_{r=1}^{M}\delta_r\mathbf{W}_{0r})\mathbf{y}_t = \sum_{t=1}^{\mathrm{T}}(\mathbf{B}_t - \bar{\mathbf{B}})^{\mathrm{T}}\mathbf{X}_t^{\mathrm{T}}\boldsymbol{\beta} + \sum_{t=1}^{\mathrm{T}}(\mathbf{B}_t - \bar{\mathbf{B}})^{\mathrm{T}}\boldsymbol{\epsilon}_t.$$

The constant $\boldsymbol{\mu}^*$ vanishes since $\sum_{t=1}^{\mathrm{T}}(\mathbf{B}_t - \bar{\mathbf{B}})^{\mathrm{T}}\boldsymbol{\mu}^* = 0$. If $\mathbf{X}_t$ is not exogenous, the operation above now weakens the correlation between $\sum_{t=1}^{T}(\mathbf{B}_t - \bar{\mathbf{B}})^{\mathrm{T}}\mathbf{X}_t^{\mathrm{T}}$ and $\sum_{t=1}^{\mathrm{T}}(\mathbf{B}_t - \bar{\mathbf{B}})^{\mathrm{T}}\boldsymbol{\epsilon}_t$, which is a standard step in estimation with instrumental variables. Consistent least squares estimation is now possible since $\boldsymbol{\beta}$ is low-dimensional as K is considered to be small. We denote the least squares estimator of $\boldsymbol{\beta}$ by

$$\begin{aligned}\boldsymbol{\beta}(\boldsymbol{\xi}, \boldsymbol{\delta}) = &\left(\sum_{t=1}^{T}\mathbf{X}_t^{\mathrm{T}}(\mathbf{B}_t - \bar{\mathbf{B}})\sum_{t=1}^{T}(\mathbf{B}_t - \bar{\mathbf{B}})^{\mathrm{T}}\mathbf{X}_t\right)^{-1} \\ &\cdot \sum_{t=1}^{T}\mathbf{X}_t^{\mathrm{T}}(\mathbf{B}_t - \bar{\mathbf{B}})\sum_{t=1}^{\mathrm{T}}(\mathbf{B}_t - \bar{\mathbf{B}})^{\mathrm{T}}\left(\mathbf{I}_N - \mathbf{A} - \sum_{r=1}^{M}\delta_r\mathbf{W}_{0r}\right)\mathbf{y}_t, \end{aligned} \tag{2.6}$$

where $\boldsymbol{\xi} = (a_{11}, \ldots, a_{1N}, \ldots, a_{N1} \ldots, a_{NN})^{\mathrm{T}}$ contains the elements of the sparse adjustment matrix in stacked notation, and $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_M)^{\mathrm{T}}$. Also define $\boldsymbol{\theta} = (\boldsymbol{\xi}^{\mathrm{T}}, \boldsymbol{\delta}^{\mathrm{T}})^{\mathrm{T}}$, so that $\boldsymbol{\theta}$ is a column vector

of all elements of $\mathbf{A}$ and all δ_r's.

Given the profiled $\boldsymbol{\beta}(\boldsymbol{\theta})$, we proceed to the estimation of $\mathbf{A}^*$. Since the length of $\boldsymbol{\theta}$ is N^2+M which can be larger than the sample size, the least squares problem can be ill-defined. Ridge regression is a traditional way to circumvent this problem. Introducing a penalty term $\|\boldsymbol{\xi}\|$ in the estimation problem effectively restricts the magnitude of $\boldsymbol{\xi}$, where the notation $\|\cdot\|$ is the L_2-norm of a vector. The estimator will be non-zero everywhere in general. Tibshirani (1996) discovers that if the penalty is $\|\cdot\|_1$ instead of $\|\cdot\|$ in a classical regression model, where $\|\mathbf{v}\|_1 = \sum_i |v_i|$ for a vector $\mathbf{v} = (v_i)$, then the estimator has elements estimated at exactly 0, thus achieving variable selection. This also suits our purpose, since $\boldsymbol{\xi}^*$ is assumed sparse. In view of this, we propose to apply the LASSO to estimate $\boldsymbol{\xi}$.

To accommodate instrumental variables in the LASSO framework, we define the following quantities. Let $\widetilde{\mathbf{y}}_i$ and $\widetilde{\mathbf{X}}_i$ be the outcome and covariates filtered through the instrumental variables,

$$\widetilde{\mathbf{y}}_i \;=\; \sum_{t=1}^{T}(\mathbf{b}_{t,i}-\bar{\mathbf{b}}_i)^{\mathrm{T}}\boldsymbol{\gamma}\mathbf{y}_t \quad \text{and} \quad \widetilde{\mathbf{X}}_i \;=\; \sum_{t=1}^{T}(\mathbf{b}_{t,i}-\bar{\mathbf{b}}_i)^{\mathrm{T}}\boldsymbol{\gamma}\mathbf{X}_t, \quad i=1,\ldots,N,$$

where $\mathbf{b}_{t,i}^{\mathrm{T}}$ is the ith row of $\mathbf{B}_t$, and $\bar{\mathbf{b}}_i^{\mathrm{T}}$ is the ith row of $\bar{\mathbf{B}}$. Thus $\widetilde{\mathbf{y}}_i$ and $\widetilde{\mathbf{X}}_i$ are of dimensions $N\times 1$ and $N\times K$, respectively. The LASSO problem is then

$$\begin{aligned}\widetilde{\theta} \;&=\; \arg\min_{\theta}\; \frac{1}{2T}\sum_{i=1}^{N}\Big\|(\mathbf{I}_N-\mathbf{A}-\sum_{r=1}^{M}\delta_r\mathbf{W}_{0r})\widetilde{\mathbf{y}}_i-\widetilde{\mathbf{X}}_i\boldsymbol{\beta}(\theta)\Big\|^2+\lambda_T\|\boldsymbol{\xi}\|_1, \qquad (2.7)\\ &\qquad \text{subj. to } \Big|\Big(\mathbf{A}+\sum_{r=1}^{M}\delta_r\mathbf{W}_{0r}\Big)\mathbf{1}_N\Big|<\mathbf{1}_N \quad \text{with} \quad |\boldsymbol{\delta}^{\mathrm{T}}\mathbf{1}_M|\le 1.\end{aligned}$$

The normalization $1/(2T)$ facilitates proofs of all theorems, and λ_T is a tuning parameter controlling the magnitude of $\boldsymbol{\xi}$ and the number of non-zeros within. We do not penalize $\boldsymbol{\delta}$ at this stage, since the analysis of the theoretical properties of $\widetilde{\boldsymbol{\xi}}$ is facilitated without penalizing both $\boldsymbol{\xi}$ and $\boldsymbol{\delta}$ at the same rate.

We do not use the maximum likelihood approach in our setting, since we do not assume normality of the residual, which is allowed to have serial dependence as defined in Section 3.3. Compared to the QMLE approach of Lee and Yu (2010), our need for both T and N to diverge arises from the fact that the spatial weight matrices are parameters to be estimated, rather than fixed. Considering them as parameters increases the difficulty substantially, since the vector of unknown is of order N^2. Still, in some settings in our paper N can grow faster than T. See Remark 2 in Section 3.4. This also explains why we do not use the generalized method of moments (GMM) since it is also difficult to implement under such a scenario. A penalized QMLE or GMM approach could be possible, and we leave this for future research.

In Section 4, we present a modified block coordinate descent (BCD) algorithm adapted to solve

(2.7). With $\widetilde{\boldsymbol{\theta}}$, the LASSO estimator $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta}(\widetilde{\boldsymbol{\theta}})$ becomes an estimator of $\boldsymbol{\beta}^*$. The tuning parameter λ_T is specified in Section 3.3 when we introduce the assumptions for our theoretical results.

If $\mathbf{W}_{0r} = \mathbf{0}$ for $r = 1, \ldots, M$, meaning that no specifications are available, there is no need to estimate $\boldsymbol{\delta}$, and problem (2.7) is still well-defined since $\mathbf{W}_{0r} = \mathbf{0}$. Thus, $\boldsymbol{\delta}$ disappears completely.

2.3 Penalized estimators - adaptive LASSO

In classical linear regression, in order for LASSO estimators to be sparse and to enjoy sign consistency (i.e., zeros are estimated as exactly zeros, non-zeros are estimated with correct signs), a stringent condition called the "irrepresentable condition" must be satisfied. Zhao and Yu (2006) provide more details. This condition arises from the fact that the penalty term $\lambda_T \|\boldsymbol{\xi}\|_1$ penalizes each element in $\boldsymbol{\xi}$ under the same tuning parameter λ_T. With a larger λ_T, small elements in $\boldsymbol{\xi}$ are driven to 0, but large elements then receive excessive penalization. If λ_T is smaller, large elements are penalized less, but small elements may not be driven exactly to 0.

This problem is resolved in Zou (2006) elegantly by the use of a penalty

$$\lambda_T \mathbf{v}^{\mathrm{T}} |\boldsymbol{\xi}| = \sum_i \frac{\lambda_T}{|\check{\xi}_i|} \xi_i,$$

where $\check{\boldsymbol{\xi}} = (\check{\xi}_i)$ is an initial estimator for $\boldsymbol{\xi}^*$. If the number of parameters to be estimated is smaller than the sample size, Zou (2006) suggests using the least squares estimator of $\boldsymbol{\xi}^*$ as an initial estimator. We choose the LASSO estimator $\widetilde{\boldsymbol{\xi}}$ for this purpose since our sample size in (2.7) is effectively N^2 but we have $N^2 + M$ parameters to estimate. Now each ξ_i is penalized by a different tuning parameter $\lambda_T/|\widetilde{\xi}_i|$. If ξ_i^* is small, then the LASSO estimator $\widetilde{\xi}_i$ should be 0 or very small, so that $\lambda_T/|\widetilde{\xi}_i|$ is positive infinity or very large, which should drive the estimator of ξ_i^* to be exactly 0. On the other hand, if ξ_i^* is large, then $\widetilde{\xi}_i$ should be large, so that $\lambda_T/|\widetilde{\xi}_i|$ is small, and the bias incurred in the estimator of ξ_i^* is thus reduced. The method has an adaptive tuning parameter for each variable, hence the name "adaptive LASSO". For classical linear regression, estimated regression coefficients under adaptive LASSO are sign-consistent, with non-zero estimators asymptotically normal and unbiased, and there is no need for the stringent irrepresentable condition.

2.3.1 When specifications $\mathbf{W}_{01}, \ldots, \mathbf{W}_{0M}$ exist

Motivated by the issues described above, when specifications $\mathbf{W}_{01}, \ldots, \mathbf{W}_{0M}$ exist, we consider the adaptive LASSO problem

$$\begin{aligned} \widehat{\boldsymbol{\xi}} &= \underset{\boldsymbol{\xi}}{\arg\min}\ \frac{1}{2T}\sum_{i=1}^{N}\Big\|(\mathbf{I}_N - \mathbf{A} - \sum_{r=1}^{M}\widetilde{\delta}_r\mathbf{W}_{0r})\widetilde{\mathbf{y}}_i - \widetilde{\mathbf{X}}_i\boldsymbol{\beta}(\widetilde{\theta})\Big\|^2 + \lambda_T\mathbf{v}^{\mathrm{T}}|\boldsymbol{\xi}|, \qquad (2.8)\\ &\quad \text{subj. to } \Big|\Big(\mathbf{A} + \sum_{r=1}^{M}\widetilde{\delta}_r\mathbf{W}_{0r}\Big)\mathbf{1}_N\Big| < \mathbf{1}_N. \end{aligned}$$

where $\mathbf{v} = (|\widetilde{\xi}_1|^{-1}, \ldots, |\widetilde{\xi}_{N^2}|^{-1}|)^{\mathrm{T}}$, $|\boldsymbol{\xi}| = (|\xi_1|, \ldots, |\xi_{N^2}|)^{\mathrm{T}}$. We replace the variables $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ by their respective LASSO estimators $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta}(\widetilde{\boldsymbol{\theta}})$ and $\widetilde{\boldsymbol{\delta}}$ from solving (2.7), so that the above equation becomes a proper adaptive LASSO problem for a regression-like setup. It can be solved by one more step of the modified BCD algorithm, to be introduced in Section 4. The tuning parameter λ_T stays the same as in the LASSO problem (2.7). The sparse adjustment matrix $\widehat{\mathbf{A}}$ is then constructed back from $\widehat{\boldsymbol{\xi}}$. The adaptive LASSO estimator for $\boldsymbol{\beta}^*$ is $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}(\widehat{\boldsymbol{\xi}}, \widetilde{\boldsymbol{\delta}})$.

For selection purpose, we want to find a sparse estimator of $\boldsymbol{\delta}^*$ as well. The estimator indicates which specified spatial weight matrix most contributes to the spatial interaction patterns observed from the data. To this end, we propose the following:

$$\begin{aligned} \widehat{\boldsymbol{\delta}} &= \underset{\boldsymbol{\delta}}{\arg\min}\ \frac{1}{2T}\sum_{i=1}^{N}\Big\|(\mathbf{I}_N - \widehat{\mathbf{A}} - \sum_{r=1}^{M}\delta_r\mathbf{W}_{0r})\widetilde{\mathbf{y}}_i - \widetilde{\mathbf{X}}_i\boldsymbol{\beta}(\widehat{\boldsymbol{\xi}}, \boldsymbol{\delta})\Big\|^2 + \lambda_T'\mathbf{u}^{\mathrm{T}}|\boldsymbol{\delta}|, \qquad (2.9)\\ &\quad \text{subj. to } \Big|\Big(\widehat{\mathbf{A}} + \sum_{r=1}^{M}\delta_r\mathbf{W}_{0r}\Big)\mathbf{1}_N\Big| < \mathbf{1}_N \quad \text{with} \quad |\boldsymbol{\delta}^{\mathrm{T}}\mathbf{1}_M| \le 1. \end{aligned}$$

where $\mathbf{u} = (|\widetilde{\delta}_1|^{-1}, \ldots, |\widetilde{\delta}_M|^{-1}|)^{\mathrm{T}}$ and $|\boldsymbol{\delta}| = (|\delta_1|, \ldots, |\delta_M|)^{\mathrm{T}}$. The function $\boldsymbol{\beta}(\widehat{\boldsymbol{\xi}}, \boldsymbol{\delta})$ is defined in (2.6), with $\boldsymbol{\xi}$ replaced by the estimator $\widehat{\boldsymbol{\xi}}$. The tuning parameter λ_T' is different from λ_T in general, and its choice is discussed in Section 4. However, these two parameters grow at the same rate, as stated in Assumption R7 in Section 3. The estimated spatial autoregressive parameter is then

$$\widehat{\rho} = \sum_{i=1}^{M}\widehat{\delta}_i.$$

Theorem 5 presents the sign-consistency and asymptotic normality of the $\widehat{\boldsymbol{\delta}}$ under certain conditions. In practice, the difference in the performance between $\boldsymbol{\beta}(\widehat{\boldsymbol{\xi}}, \widetilde{\boldsymbol{\delta}})$ and $\boldsymbol{\beta}(\widehat{\boldsymbol{\xi}}, \widehat{\boldsymbol{\delta}})$ is negligible, so we set $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}(\widehat{\boldsymbol{\xi}}, \widetilde{\boldsymbol{\delta}})$.

2.3.2 When specifications $\mathbf{W}_{01}, \ldots, \mathbf{W}_{0M}$ do not exist

If no specifications are available, then $\mathbf{A}^*$ must act as the spatial weight matrix itself. The adaptive LASSO problem is to solve

$$\begin{aligned} \widehat{\boldsymbol{\xi}} &= \arg\min_{\boldsymbol{\xi}} \frac{1}{2T}\sum_{i=1}^{N}\left\|(\mathbf{I}_N-\mathbf{A})\widetilde{\mathbf{y}}_i-\widetilde{\mathbf{X}}_i\boldsymbol{\beta}(\boldsymbol{\xi})\right\|^2+\lambda_T\mathbf{v}^{\mathrm{T}}|\boldsymbol{\xi}|, \\ &\quad \text{subj. to } |\mathbf{A}\mathbf{1}_N|<\mathbf{1}_N \quad \text{with} \quad |\boldsymbol{\delta}^{\mathrm{T}}\mathbf{1}_M|\le 1. \end{aligned} \tag{2.10}$$

The difference between (2.8) and (2.10) is that $\boldsymbol{\beta}$ is set at $\widetilde{\boldsymbol{\beta}}=\boldsymbol{\beta}(\widetilde{\boldsymbol{\theta}})$, the LASSO estimator, for (2.8), while $\boldsymbol{\beta}$ is set at $\boldsymbol{\beta}(\boldsymbol{\xi})$ for (2.10), so that the problem above is not exactly a penalized regression setup. The main reason we need to solve (2.10) rather than (2.8) is that when specifications do not exist, $\mathbf{A}^*$ is a proper spatial weight matrix, and so has at least order N number of non-zero elements (e.g., assuming only 1 non-zero in each row of $\mathbf{A}^*$, then $\mathbf{A}^*$ has exactly N non-zero elements). Assumption M2 is no longer valid, but is replaced by M2' in Section 3.3. The proof of partial sign-consistency and asymptotic normality of $\widehat{\boldsymbol{\xi}}$ is also changed slightly, compared to the proof of Theorem 3 with specifications. A solution from (2.10) greatly facilitates the proof. With $\widehat{\boldsymbol{\xi}}$, the adaptive LASSO estimator for $\boldsymbol{\beta}^*$ is $\widehat{\boldsymbol{\beta}}=\boldsymbol{\beta}(\widehat{\boldsymbol{\xi}})$.

Remark 1. We penalize on $\boldsymbol{\xi}$ and $\boldsymbol{\delta}$ in two different problems in Section 2.3.1. The reasons we do this are twofold. First, penalizing on both $\boldsymbol{\xi}$ and $\boldsymbol{\delta}$ at the same time makes the proof of theoretical results more difficult. The dimension of $\boldsymbol{\xi}$ is N^2 which grows with T, while that for $\boldsymbol{\delta}$ is M. This dimension, M, relies on the knowledge of the researcher. The tuning parameters needed for $\boldsymbol{\xi}$ and $\boldsymbol{\delta}$ can potentially be very different. Second, penalizing on both of them first results in two LASSO estimators, which are not as accurate as the adaptive LASSO estimators. Hence, a second penalization is needed, on both $\boldsymbol{\xi}$ and $\boldsymbol{\delta}$. The computational complexities of finding the tuning parameters of two problems, penalizing on two parameters, will be more onerous than finding the tuning parameter of penalizing $\boldsymbol{\xi}$ first, and then on $\boldsymbol{\xi}$ and $\boldsymbol{\delta}$ separately.

3 Properties of the Estimators

3.1 Full matrix notations for the LASSO and adaptive LASSO problems

In Section 2 we present the main LASSO and adaptive LASSO problems. They can actually be compactly presented in matrix notations, which facilitates the presentation of our theoretical results. We first introduce the notation

$$\mathbf{B}=T^{-1/2}N^{-a/2}(\mathbf{B}_{\boldsymbol{\gamma}}-\bar{\mathbf{B}}_{\boldsymbol{\gamma}})=T^{-1/2}N^{-a/2}\mathbf{I}_N\otimes\{(\mathbf{I}_N\otimes\boldsymbol{\gamma}^{\mathrm{T}})(\mathbf{B}_1-\bar{\mathbf{B}},\ldots,\mathbf{B}_T-\bar{\mathbf{B}})^{\mathrm{T}}\}. \tag{3.1}$$

The constant a is to be introduced in Assumption (R4) in Section 3.3. In general, a larger a means the exogenous variables in $\mathbf{B}_t$ correlates with more covariates in $\mathbf{X}_t$. In practice we do not know

a, and we can calculate $\mathbf{B}$ by setting $a = 1$. This does not change the optimal values of any tuning parameters and estimators in all LASSO and adaptive LASSO problems presented using $\mathbf{B}$ below.

We rewrite (2.4) as

$$\mathbf{y} = \boldsymbol{\mu}^* \otimes \mathbf{1}_T + \mathbf{Z}\boldsymbol{\xi}^* + \mathbf{Z}\mathbf{V}_0\boldsymbol{\delta}^* + \mathbf{X}_{\boldsymbol{\beta}^*}\mathrm{vec}(\mathbf{I}_N) + \boldsymbol{\epsilon}, \tag{3.2}$$

where $\mathbf{y} = \mathrm{vec}\{(\mathbf{y}_1, \ldots, \mathbf{y}_T)^{\mathrm{T}}\}$, $\boldsymbol{\epsilon} = \mathrm{vec}\{(\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_T)^{\mathrm{T}}\}$, $\mathbf{Z} = \mathbf{I}_N \otimes (\mathbf{y}_1, \ldots, \mathbf{y}_T)^{\mathrm{T}}$, $\boldsymbol{\delta}^* = (\delta_1^*, \ldots, \delta_M^*)^{\mathrm{T}}$, $\mathbf{X}_{\boldsymbol{\beta}^*} = \mathbf{I}_N \otimes \{(\mathbf{I}_T \otimes \boldsymbol{\beta}^{*\mathrm{T}})(\mathbf{X}_1, \ldots, \mathbf{X}_T)^{\mathrm{T}}\}$, $\boldsymbol{\xi}^* = \mathrm{vec}(\mathbf{A}^{*\mathrm{T}})$, and $\mathbf{V}_0 = (\mathrm{vec}(\mathbf{W}_{01}^{\mathrm{T}}), \ldots, \mathrm{vec}(\mathbf{W}_{0M}^{\mathrm{T}}))$. The notation $\otimes$ represents the Kronecker product. Finally, $\mathrm{vec}(\cdot)$ is the vectorization operator for a matrix, column by column. The model now has design matrices $\mathbf{Z}$ and $\mathbf{Z}\mathbf{V}_0$ in a classical linear regression setting, with $\boldsymbol{\xi}^*$ and $\boldsymbol{\delta}^*$ being the true regression parameters, except for the fact that $\mathbf{Z}$ contains the endogenous variables $\mathbf{y}_t$. Multiplying both sides by $\mathbf{B}^{\mathrm{T}}$, we then have an augmented model:

$$\mathbf{B}^{\mathrm{T}}\mathbf{y} = \mathbf{B}^{\mathrm{T}}\mathbf{Z}(\boldsymbol{\xi}^* + \mathbf{V}_0\boldsymbol{\delta}^*) + \mathbf{B}^{\mathrm{T}}\mathbf{X}_{\boldsymbol{\beta}^*}\mathrm{vec}(\mathbf{I}_N) + \mathbf{B}^{\mathrm{T}}\boldsymbol{\epsilon}. \tag{3.3}$$

To express the LASSO problem in matrix form, we first define $\mathbf{y}^v = (\mathbf{y}_1^{\mathrm{T}}, \ldots, \mathbf{y}_T^{\mathrm{T}})^{\mathrm{T}}$, so that model (2.4) becomes

$$\mathbf{y}^v = \mathbf{1}_T \otimes \boldsymbol{\mu}^* + \Big(\mathbf{A}^{*\otimes} + \sum_{i=1}^{M} \delta_i^* \mathbf{W}_{0i}^{\otimes}\Big)\mathbf{y}^v + \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}^v,$$

where $\boldsymbol{\epsilon}^v$ is defined similar to $\mathbf{y}^v$, $\mathbf{X} = (\mathbf{X}_1^{\mathrm{T}}, \ldots, \mathbf{X}_T^{\mathrm{T}})^{\mathrm{T}}$, and for any matrix C, $C^{\otimes} = \mathbf{I}_T \otimes C$. Then $\boldsymbol{\beta}(\boldsymbol{\theta})$ defined in (2.6) can be expressed as

$$\boldsymbol{\beta}(\boldsymbol{\theta}) = (\mathbf{X}^{\mathrm{T}}\mathbf{B}^v\mathbf{B}^{v\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{B}^v\mathbf{B}^{v\mathrm{T}}\Big(\mathbf{I}_{TN} - \mathbf{A}^{\otimes} - \sum_{i=1}^{M} \delta_i \mathbf{W}_{0i}^{\otimes}\Big)\mathbf{y}^v.$$

Using (3.3), the LASSO problem (2.7) can then be rewritten as

$$\begin{aligned} \widetilde{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} & \frac{1}{2T}\big\|\mathbf{B}^{\mathrm{T}}\mathbf{y} - \mathbf{B}^{\mathrm{T}}\mathbf{Z}\boldsymbol{\xi} - \mathbf{B}^{\mathrm{T}}\mathbf{Z}\mathbf{V}_0\boldsymbol{\delta} - \mathbf{B}^{\mathrm{T}}\mathbf{X}_{\boldsymbol{\beta}(\boldsymbol{\theta})}\mathrm{vec}(\mathbf{I}_N)\big\|^2 + \lambda_T\|\boldsymbol{\xi}\|_1, \\ \text{subj. to } & \Big|\Big(\mathbf{A} + \sum_{r=1}^{M} \delta_r \mathbf{W}_{0r}\Big)\mathbf{1}_N\Big| < \mathbf{1}_N \quad \text{with} \quad |\boldsymbol{\delta}^{\mathrm{T}}\mathbf{1}_M| \le 1. \end{aligned} \tag{3.4}$$

The formula for $\widetilde{\boldsymbol{\delta}}$ for the LASSO problem above is

$$\begin{aligned}
\widetilde{\boldsymbol{\delta}} &= \Big[(\mathbf{H}-\mathbf{B}^{\mathrm{T}}\mathbf{Z}\mathbf{V}_0)^{\mathrm{T}}(\mathbf{H}-\mathbf{B}^{\mathrm{T}}\mathbf{Z}\mathbf{V}_0)\Big]^{-1}(\mathbf{H}-\mathbf{B}^{\mathrm{T}}\mathbf{Z}\mathbf{V}_0)^{\mathrm{T}} \\
&\quad \cdot (\mathbf{B}^{\mathrm{T}}\mathbf{Z}\widetilde{\boldsymbol{\xi}}-\mathbf{B}^{\mathrm{T}}\mathbf{y}+\mathbf{K}(\mathbf{I}_{TN}-\widetilde{\mathbf{A}}^{\otimes})\mathbf{y}^{v}), \text{ with} \\
\mathbf{K} &= T^{-1/2}N^{-a/2}\Big(\sum_{t=1}^{T}\mathbf{X}_t\otimes(\mathbf{B}_t-\bar{\mathbf{B}})\boldsymbol{\gamma}\Big)(\mathbf{X}^{\mathrm{T}}\mathbf{B}^{v}\mathbf{B}^{v\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{B}^{v}\mathbf{B}^{v\mathrm{T}}, \\
\mathbf{H} &= \mathbf{K}(\mathbf{W}_{01}^{\otimes}\cdots\mathbf{W}_{0M}^{\otimes})(\mathbf{I}_M\otimes\mathbf{y}^{v}).
\end{aligned} \tag{3.5}$$

The term $N^{-a/2}$ is cancelled out in the formula for $\widetilde{\boldsymbol{\delta}}$ above. The adaptive LASSO problem (2.8) with $M>0$, can be rewritten as

$$\begin{aligned}
\widehat{\boldsymbol{\xi}} &= \arg\min_{\boldsymbol{\xi}}\frac{1}{2T}\big\|\mathbf{B}^{\mathrm{T}}\mathbf{y}-\mathbf{B}^{\mathrm{T}}\mathbf{Z}\boldsymbol{\xi}-\mathbf{B}^{\mathrm{T}}\mathbf{Z}\mathbf{V}_0\widetilde{\boldsymbol{\delta}}-\mathbf{B}^{\mathrm{T}}\mathbf{X}_{\widetilde{\boldsymbol{\beta}}}\mathrm{vec}(\mathbf{I}_N)\big\|^2+\lambda_T\mathbf{v}^{\mathrm{T}}|\boldsymbol{\xi}|, \\
&\quad \text{subj. to } \Big|\Big(\mathbf{A}+\sum_{r=1}^{M}\widetilde{\delta}_r\mathbf{W}_{0r}\Big)\mathbf{1}_N\Big|<\mathbf{1}_N.
\end{aligned} \tag{3.6}$$

Finally, the adaptive LASSO problem for selecting $\boldsymbol{\delta}$ in (2.9) can be rewritten as

$$\begin{aligned}
\widehat{\boldsymbol{\delta}} &= \arg\min_{\boldsymbol{\delta}}\frac{1}{2T}\big\|\mathbf{B}^{\mathrm{T}}\mathbf{y}-\mathbf{B}^{\mathrm{T}}\mathbf{Z}\widehat{\boldsymbol{\xi}}-\mathbf{B}^{\mathrm{T}}\mathbf{Z}\mathbf{V}_0\boldsymbol{\delta}-\mathbf{B}^{\mathrm{T}}\mathbf{X}_{\boldsymbol{\beta}(\widehat{\boldsymbol{\xi}},\boldsymbol{\delta})}\mathrm{vec}(\mathbf{I}_N)\big\|^2+\lambda'_T\mathbf{u}^{\mathrm{T}}|\boldsymbol{\delta}|, \\
&\quad \text{subj. to } \Big|\Big(\widehat{\mathbf{A}}+\sum_{r=1}^{M}\delta_r\mathbf{W}_{0r}\Big)\mathbf{1}_N\Big|<\mathbf{1}_N \quad \text{with} \quad |\boldsymbol{\delta}^{\mathrm{T}}\mathbf{1}_M|\le 1.
\end{aligned} \tag{3.7}$$

A more direct penalized least squares formulation exist for this problem

$$\begin{aligned}
\widehat{\boldsymbol{\delta}} &= \arg\min_{\boldsymbol{\delta}}\frac{1}{2T}\big\|\mathbf{B}^{\mathrm{T}}\mathbf{y}-\mathbf{B}^{\mathrm{T}}\mathbf{Z}\widehat{\boldsymbol{\xi}}-\widehat{\mathbf{h}}-(\mathbf{B}^{\mathrm{T}}\mathbf{Z}\mathbf{V}_0-\mathbf{H})\boldsymbol{\delta}\big\|^2+\lambda'_T\mathbf{u}^{\mathrm{T}}|\boldsymbol{\delta}|, \text{ with} \\
\widehat{\mathbf{h}} &= T^{-1/2}N^{-a/2}\Big(\sum_{t=1}^{T}\mathbf{X}_t\otimes(\mathbf{B}_t-\bar{\mathbf{B}})\boldsymbol{\gamma}\Big)(\mathbf{X}^{\mathrm{T}}\mathbf{B}^{v}\mathbf{B}^{v\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{B}^{v}\mathbf{B}^{v\mathrm{T}}(\mathbf{I}_{TN}-\widehat{\mathbf{A}}^{\otimes})\mathbf{y}^{v}, \\
&\quad \text{subj. to } \Big|\Big(\widehat{\mathbf{A}}+\sum_{r=1}^{M}\delta_r\mathbf{W}_{0r}\Big)\mathbf{1}_N\Big|<\mathbf{1}_N \quad \text{with} \quad |\boldsymbol{\delta}^{\mathrm{T}}\mathbf{1}_M|\le 1.
\end{aligned} \tag{3.8}$$

When $M=0$, problem (2.10) can be rewritten as

$$\widehat{\boldsymbol{\xi}} = \arg\min_{\boldsymbol{\xi}}\frac{1}{2T}\big\|\mathbf{B}^{\mathrm{T}}\mathbf{y}-\mathbf{B}^{\mathrm{T}}\mathbf{Z}\boldsymbol{\xi}-\mathbf{B}^{\mathrm{T}}\mathbf{X}_{\boldsymbol{\beta}(\boldsymbol{\xi})}\mathrm{vec}(\mathbf{I}_N)\big\|^2+\lambda_T\mathbf{v}^{\mathrm{T}}|\boldsymbol{\xi}|, \text{ subj. to } |\mathbf{A}\mathbf{1}_N|<\mathbf{1}_N. \tag{3.9}$$

An equivalent problem which is a proper penalized regression setup is given by

$$\begin{aligned}\widehat{\boldsymbol{\xi}} &= \arg\min_{\boldsymbol{\xi}} \frac{1}{2T}\big\|\mathbf{B}^{\mathrm{T}}\mathbf{y} - (\mathbf{B}^{\mathrm{T}}\mathbf{Z} - \mathbf{K}')\boldsymbol{\xi} - \mathbf{K}\mathbf{y}^{v}\big\|^2 + \lambda_T \mathbf{v}^{\mathrm{T}}|\boldsymbol{\xi}|, \text{ subj. to } |\mathbf{A}\mathbf{1}_N| < \mathbf{1}_N, \text{ with} \\ \mathbf{K}' &= T^{-1/2}N^{-a/2}\Big(\sum_{t=1}^{T}\mathbf{X}_t \otimes (\mathbf{B}_t - \bar{\mathbf{B}})\boldsymbol{\gamma}\Big)(\mathbf{X}^{\mathrm{T}}\mathbf{B}^{v}\mathbf{B}^{v\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{B}^{v}\Big(\sum_{t=1}^{T}(\mathbf{B}_t - \bar{\mathbf{B}})^{\mathrm{T}} \otimes \mathbf{y}_t^{\mathrm{T}}\Big).\end{aligned} \tag{3.10}$$

3.2 Time series variables and notations

We present notations used hereinafter, and introduce the measure of time dependence of all the time series variables involved. The concept of approximate sparsity of the adjustment matrix $\mathbf{A}^*$ in (2.3) is introduced in Section 3.3, together with all assumptions we use in this paper. Theorems are presented in Section 3.4.

Denote $\{\mathbf{b}_t\} = \{\mathrm{vec}(\mathbf{B}_t)\}$ and $\mathbf{x}_t = \{\mathrm{vec}(\mathbf{X}_t)\}$ the vectorized processes for $\{\mathbf{B}_t\}$ and $\{\mathbf{X}_t\}$, with length NL and NK, respectively. For $t = 1, \ldots, T$, we assume that

$$\mathbf{x}_t = [f_j(\mathcal{F}_t)]_{1\le j\le NK}, \quad \mathbf{b}_t = [g_j(\mathcal{G}_t)]_{1\le j\le NL}, \quad \boldsymbol{\epsilon}_t = [h_j(\mathcal{H}_t)]_{1\le j\le N}, \tag{3.11}$$

where the $f_j(\cdot)$'s, $g_j(\cdot)$'s and $h_j(\cdot)$'s are measurable functions defined on the real line, and $\mathcal{F}_t = (\ldots, \mathbf{e}_{x,t-1}, \mathbf{e}_{x,t})$, $\mathcal{G}_t = (\ldots, \mathbf{e}_{b,t-1}, \mathbf{e}_{b,t})$, $\mathcal{H}_t = (\ldots, \mathbf{e}_{\epsilon,t-1}, \mathbf{e}_{\epsilon,t})$ are defined by independent and identically distributed processes $\{\mathbf{e}_{x,t}\}$ $\{\mathbf{e}_{b,t}\}$ and $\{\mathbf{e}_{\epsilon,t}\}$ respectively, with $\{\mathbf{e}_{b,t}\}$ independent of $\{\mathbf{e}_{\epsilon,t}\}$ but correlated with $\{\mathbf{e}_{x,t}\}$.

We use the functional dependence measure introduced by Wu (2005) for gauging the serial dependence of a process. For $d > 0$, denoting $\mathbf{x}_t = (x_{it}), \mathbf{b}_t = (b_{jt})$ and $\boldsymbol{\epsilon}_t = (\epsilon_{\ell t})$, we define

$$\begin{aligned}\theta^{x}_{t,d,i} &= \big\|x_{it} - x'_{it}\big\|_d = (E|x_{it} - x'_{it}|^d)^{1/d}, \quad i = 1, \ldots, NK, \\ \theta^{b}_{t,d,j} &= \big\|b_{jt} - b'_{jt}\big\|_d = (E|b_{jt} - b'_{jt}|^d)^{1/d}, \quad j = 1, \ldots, NL, \\ \theta^{\epsilon}_{t,d,\ell} &= \big\|\epsilon_{\ell t} - \epsilon'_{\ell} t\big\|_d = (E|\epsilon_{\ell t} - \epsilon'_{\ell t}|^d)^{1/d}, \quad \ell = 1, \ldots, N,\end{aligned} \tag{3.12}$$

where $x'_{it} = f_i(\mathcal{F}'_t)$, $\mathcal{F}'_t = (\ldots, \mathbf{e}_{x,-1}, \mathbf{e}'_{x,0}, \mathbf{e}_{x,1}, \ldots, \mathbf{e}_{x,t})$, with $\mathbf{e}'_{x,0}$ independent of all other $\mathbf{e}_{x,j}$'s. Hence x'_{it} is a coupled version of x_{it} with $\mathbf{e}_{x,0}$ replaced by an i.i.d. copy $\mathbf{e}'_{x,0}$. Intuitively, a large $\theta^{x}_{t,d,i}$ means that serial correlation is strong, at least for variables at most time t apart. Finally, we have similar definitions for b'_{jt} and $\epsilon'_{\ell t}$. Such a definition of "physical" or functional dependence of time series on past "inputs" is used by various papers, for example Shao (2010) and Zhou (2010).

3.3 Assumptions and partial sign consistency

Assumptions in this paper are effective whether there are specifications $\mathbf{W}_{01}, \ldots, \mathbf{W}_{0M}$ (so that $M > 0$) or not (so that $M = 0$). The exceptions are M2, R5 and R8, which apply to the former scenario while M2', R5' and R8' are for the latter. These assumptions concern the elements in the sparse adjustment matrix $\mathbf{A}^*$, the potential specifications $\mathbf{W}_{0i}$, the true spatial weight

matrix $\mathbf{W}^*$ in (2.3), and the distributional behavior and serial dependence of all the time series variables involved. The identification condition of model (3.3) is also presented as Assumption M5, with the identification of model (3.3) proved in Section 3.3.2. The partial sign consistency of an estimator of an approximately sparse matrix is discussed when we explain Assumptions M2 and M2'. Assumptions that start with "R" are in general more of a technical nature than those starting with "M".

M1. With the true spatial weight matrix $\mathbf{W}^*$ defined in (2.3), there exists a constant $\eta > 0$ such that $\left\|\mathbf{W}^*\right\|_\infty < \eta < 1$ uniformly as $N \to \infty$. The elements in $\mathbf{W}^*$ can be negative, and $\mathbf{W}^*$ can be asymmetric.

M2. (Approximate sparseness for $\mathbf{A}^*$ when $M > 0$) There exists a constant $\tau > 0$ such that the elements a_{ij}^* of $\mathbf{A}^*$ are constants as $N \to \infty$ whenever they are larger than or equal to τ in magnitude. For those elements smaller than τ, we have either $a_{ij}^* = 0$ or $a_{ij}^* \to 0$ as $N \to \infty$.

Define

$$\begin{aligned} &J_0 = \{j : \xi_j^* = 0 \text{ and not corr. to the diagonal of } \mathbf{A}^*\}, \\ &J_1 = \{j : |\xi_j^*| \geq \tau\},\ J_2 = \{j : 0 < |\xi_j^*| < \tau\}. \end{aligned} \tag{3.13}$$

Denote $n = |J_1|$. Then the number of elements belonging to J_1 in each row of $\mathbf{A}^*$ is bounded uniformly away from infinity as $N \to \infty$. Moreover, $n = o(N^{1/2})$.

M2' (Approximate sparseness for $\mathbf{A}^*$ when $M = 0$) Same as M2, except that $n = O(N)$.

M3. The processes $\{\mathbf{B}_t\}$, $\{\mathbf{X}_t\}$ and $\{\boldsymbol{\epsilon}_t\}$ defined in Section 2 are second-order stationary, with $\{\mathbf{X}_t\}$ and $\{\boldsymbol{\epsilon}_t\}$ having mean zero. The exogenous variables $\{\mathbf{B}_t\}$ are independent of the noise $\{\boldsymbol{\epsilon}_t\}$. The tail condition $P(|Z| > v) \leq D_1 \exp(-D_2 v^q)$ is satisfied for the variables $B_{t,jk}, X_{t,jk}, \epsilon_{t,j}$ by the same constants D_1, D_2, and q.

M4. Define

$$\Theta_{m,a}^x = \sum_{t=m}^{\infty} \max_{1 \leq j \leq NK} \theta_{t,a,i}^x,\ \ \Theta_{m,a}^b = \sum_{t=m}^{\infty} \max_{1 \leq j \leq NL} \theta_{t,a,j}^b,\ \ \Theta_{m,a}^\epsilon = \sum_{t=m}^{\infty} \max_{1 \leq j \leq N} \theta_{t,a,\ell}^\epsilon,$$

where $\theta_{t,a,i}^x, \theta_{t,a,j}^b$ and $\theta_{t,a,\ell}^\epsilon$ are defined in (3.12).

We assume that for some $w > 2$, $\Theta_{m,2w}^x, \Theta_{m,2w}^b, \Theta_{m,2w}^\epsilon \leq Cm^{-\alpha}$ with $\alpha, C > 0$ being constants that can depend on w.

M5. (Identification condition) Assume that the two sets of parameters $(\boldsymbol{\xi}^*, \boldsymbol{\delta}^*, \boldsymbol{\beta}^*)$ and $(\boldsymbol{\xi}^o, \boldsymbol{\delta}^o, \boldsymbol{\beta}^o)$ both satisfy model (3.3). Let both $\boldsymbol{\xi}^*$ and $\boldsymbol{\xi}^o$ be exactly sparse (see also Assumption M2), with the set S defined as

$$S = \{j : \xi_j^* \neq 0 \text{ or } \xi_j^o \neq 0\}.$$

Then the identification condition is that the matrix $\mathbf{Q}^{\mathrm{T}}\mathbf{Q}$ has all its eigenvalues uniformly bounded away from 0, where

$$\mathbf{Q} = [E(\mathbf{B}^{\mathrm{T}}\mathbf{Z}_S),\ E(\mathbf{B}^{\mathrm{T}}\mathbf{Z}\mathbf{V}_0),\ E(\mathbf{B}^{\mathrm{T}}\widetilde{\mathbf{X}})] \text{ and}$$
$$\widetilde{\mathbf{X}} = (\mathbf{x}_{1,1}, \ldots, \mathbf{x}_{T,1}, \ldots, \mathbf{x}_{1,N}, \ldots, \mathbf{x}_{T,N})^{\mathrm{T}}.$$

The notation A_S means that the matrix A has columns restricted to the set S. If $M = 0$, the term $E(\mathbf{B}^{\mathrm{T}}\mathbf{Z}\mathbf{V}_0)$ is omitted.

For approximately sparse $\boldsymbol{\xi}^*$ and $\boldsymbol{\xi}^o$, we assume that those elements which are $o(1)$ are all identified exactly to 0. □

R1. Each column vector $\mathrm{vec}(\mathbf{W}_{0i}^{\mathrm{T}})$ in $\mathbf{V}_0$, $i = 1, \ldots, M$ is linearly independent of each other, such that there exists a constant $u > 0$ with $\sigma_M^2(\mathbf{V}_0) \geq u > 0$ uniformly as $N \to \infty$, where $\sigma_i(A)$ is the ith largest singular value of a matrix A.

Moreover, $\max_{1 \leq i \leq M} \left\|\mathbf{W}_{0i}\right\|_1 \leq c < \infty$ uniformly as $N \to \infty$ for some constant $c > 0$.

R2. Write $\boldsymbol{\epsilon}_t = \boldsymbol{\Sigma}_\epsilon^{1/2}\boldsymbol{\epsilon}_t^*$, where $\boldsymbol{\Sigma}_\epsilon$ is the covariance matrix for $\boldsymbol{\epsilon}_t$. Then the elements in $\boldsymbol{\Sigma}_\epsilon$ are all less than $\sigma_{\max}^2$ uniformly as $N \to \infty$. The same logic applies to the variance of the elements in $\mathbf{B}_t$.

We also assume $\left\|\boldsymbol{\Sigma}_\epsilon^{1/2}\right\|_\infty \leq S_\epsilon < \infty$ uniformly as $N \to \infty$, with $\{\epsilon_{t,j}^*\}_{1 \leq j \leq N}$ being a martingale difference with respect to the filtration generated by $\sigma(\epsilon_{t,1}^*, \ldots, \epsilon_{t,j}^*)$. The tail condition $P(|Z| > v) \leq D_1 \exp(-D_2 v^q)$ is also satisfied by $\epsilon_{t,j}^*$, while Assumption M4 is also satisfied by $\{\boldsymbol{\epsilon}_t^*\}_{1 \leq t \leq T}$.

R3. All singular values of $E(\mathbf{X}_t^{\mathrm{T}}\mathbf{B}_t)$ are uniformly larger than Nu for some constant $u > 0$, while the maximum singular value is of order N. Individual entries in the matrix $E(\mathbf{b}_t\mathbf{x}_t^{\mathrm{T}})$ are uniformly bounded away from infinity, where $\mathbf{b}_t$ and $\mathbf{x}_t$ are defined in the paragraph containing (3.11).

R4. For $a \in (0, 1)$ the same as in the definition of (3.1), define

$$\mathbf{G} = N^{-a}E(T^{-1}\mathbf{Z}^{\mathrm{T}}(\mathbf{B}_{\boldsymbol{\gamma}} - \bar{\mathbf{B}}_{\boldsymbol{\gamma}}))E(T^{-1}(\mathbf{B}_{\boldsymbol{\gamma}} - \bar{\mathbf{B}}_{\boldsymbol{\gamma}})^{\mathrm{T}}\mathbf{Z}).$$

Each block $E(T^{-1}\sum_{t=1}^T(\mathbf{B}_t - \bar{\mathbf{B}})\boldsymbol{\gamma}\mathbf{y}_t^{\mathrm{T}}) = E(T^{-1}\sum_{t=1}^T(\mathbf{B}_t - \bar{\mathbf{B}})\boldsymbol{\gamma}\boldsymbol{\beta}^{*\mathrm{T}}\mathbf{X}_t^{\mathrm{T}}\boldsymbol{\Pi}^{*\mathrm{T}})$ in the block diagonal matrix $E(T^{-1}(\mathbf{B}_{\boldsymbol{\gamma}} - \bar{\mathbf{B}}_{\boldsymbol{\gamma}})^{\mathrm{T}}\mathbf{Z})$ is assumed to have full rank N, such that there exists a constant $u > 0$ with $\lambda_{\min}(\mathbf{G}) \geq u > 0$ uniformly as $N \to \infty$. The maximum eigenvalue of $\mathbf{G}$ is uniformly bounded from infinity.

R5. For the same constant a as in Assumption R4, we have for each N,

$$\max_{1\le i\le N}\sum_{j=1}^{N}\left\|E(\mathbf{b}_{t,i}\mathbf{x}_{t,j}^{\mathrm{T}})\right\|_{\max},\quad \max_{1\le j\le N}\sum_{i=1}^{N}\left\|E(\mathbf{b}_{t,i}\mathbf{x}_{t,j}^{\mathrm{T}})\right\|_{\max}\le C_{bx}N^{a},$$

where $C_{bx}>0$ is a constant and $\mathbf{b}_{t,i}$, $\mathbf{x}_{t,j}$ are the column vectors for the ith row of $\mathbf{B}_t$ and jth row of $\mathbf{X}_t$ respectively. At the same time, $E(\mathbf{X}_t\otimes\mathbf{B}_t\boldsymbol{\gamma})$ has all singular values of order N^{1+a}.

R5' For the equation shown in R5, change the part for $E(\mathbf{X}_t\otimes\mathbf{B}_t\boldsymbol{\gamma})$ to

$$\left\|E(\mathbf{X}_t\otimes\mathbf{B}_t\boldsymbol{\gamma})\right\|_1,\ \left\|E(\mathbf{B}_t\otimes\boldsymbol{\Pi}^*\mathbf{X}_t\boldsymbol{\beta}^*)\right\|_1=O(N).$$

If the rows of the matrix $\mathbf{X}_t\otimes\mathbf{B}_t\boldsymbol{\gamma}$ or $\mathbf{B}_t\otimes\boldsymbol{\Pi}^*\mathbf{X}_t\boldsymbol{\beta}^*$ are restricted to J_1, then the above are assumed to be $o(N)$.

R6. Assume $0<b<1$. For fixed $1\le k\le K$, the eigenvalues of $\mathrm{var}(N^{-b/2}\mathbf{B}_{t,k})$ and $\mathrm{var}(\boldsymbol{\epsilon}_t)$ are uniformly bounded away from 0 and infinity, and respectively dominate the singular values from $N^{-b}E((\mathbf{B}_{t+\tau,k}-\boldsymbol{\mu}_{b,k})(\mathbf{B}_{t,k}-\boldsymbol{\mu}_{b,k})^{\mathrm{T}})$ and $E(\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_{t+\tau}^{\mathrm{T}})$. The sum of the ith largest singular values over τ for each $1\le i\le N$ is assumed finite for both $\{N^{-b/2}\mathbf{B}_t\}$ and $\{\boldsymbol{\epsilon}_t\}$.

R7. Define $c_T=dT^{-1/2}\log^{1/2}(T\vee N)$ for some constant $d>0$. The tuning parameter λ_T for (2.7) is such that $\lambda_T=Cc_T$ for some constant $C>0$. The tuning parameter λ_T' for (2.9) is $\lambda_T'=C'c_T$ for some constant $C'>0$.

R8. Assume that $M>0$. In all the statements above, we assume that as $N,T\to\infty$,

$$\begin{aligned} c_Tn,\ nN^{a-1}\log^{1/2}(T\vee N)&=o(1),\\ T^{-1/2}N^{\frac{a-b}{2}}\log(T\vee N),\ T^{-1/2}N^{\frac{b}{2}+\frac{1}{2w}}&=o(1).\end{aligned}$$

We further assume that $T^{1/2}N^{\frac{a-b}{2}}\max_{1\le j\le N}\left\|\mathbf{a}_{j,J_2}^*\right\|_1=o(1)$ and $\left\|\boldsymbol{\xi}_{J_2}^*\right\|_1=o(c_TN^{\frac{1}{2}+\frac{1}{2w}})$ which is a rate diverging to infinity.

R8' We follow R8, except that $M=0$. The consition $nN^{1-a}\log^{1/2}(T\vee N)=o(1)$ is replaced by $N^{a-b}\log(T\vee N)=o(1)$.

3.3.1 Explanations for Assumptions M1 to M4 and M2'

We now explain briefly the assumptions M1-M4 and M2'. Assumption M5 will be used in the Section 3.3.2 to prove the identification of model (3.3). Assumption M1 ensures that model (2.1) has a stable reduced form

$$\mathbf{y}_t=\boldsymbol{\Pi}^*\boldsymbol{\mu}^*+\boldsymbol{\Pi}^*\mathbf{X}_t\boldsymbol{\beta}^*+\boldsymbol{\Pi}^*\boldsymbol{\epsilon}_t,\quad \boldsymbol{\Pi}^*=(\mathbf{I}_N-\mathbf{W}^*)^{-1},\tag{3.14}$$

where the innovations $\mathbf{\Pi}^*\boldsymbol{\epsilon}_t$ have finite variances. Corrado and Fingleton (2011) uses a similar row sum condition in a slightly different spatial model specification. The row sum cannot be exactly one like traditional spatial autoregressive models with spatial weight matrix $\mathbf{W}^*$. In those models, a spatial autoregressive parameter ρ having $|\rho| < 1$ is multiplied with $\mathbf{W}^*$, and so the spatial weight matrix is in fact $\rho\mathbf{W}^*$ in those models.

Assumptions M2 and M2' are ways to relax the requirement of strict sparseness on $\mathbf{A}^*$. These are two major assumptions in this paper. The set J_0 is the set of all the zeros in the sparse adjustment matrix $\mathbf{A}^*$ excluding the diagonal elements, which are all 0 by definition. Both J_1 and J_2 are the sets for the non-zeros, but those elements in J_2, according to assumption M2, are all $o(1)$ as both $T, N \to \infty$. A partially sign consistent estimator will estimate all the elements in J_0 and J_2 to be exactly 0, while estimating those in J_1 to be non-zero with the correct sign. Such a regularized estimator can accumulate smaller errors than those allowing for non-zero estimates for the small entries, since the estimation errors involved in estimating all those small entries can be much larger than setting them to zero. Theorem 3 shows that $\widehat{\boldsymbol{\xi}}$ in (3.6) and (3.10) for $M > 0$ and $M = 0$, respectively, are partially sign-consistent with probability going to 1.

Assumption M2 states that $n = o(N^{1/2})$. This means that when $M > 0$, the best linear combination of the specified spatial weight matrices has only $o(N^{1/2})$ number of non-zero elements not yet close enough to the corresponding true values. With reasonable specified spatial weight matrices, this is usually true. However, when $M = 0$, Assumption M2' has $n = O(N)$, reflecting that each row of the spatial weight matrix $\mathbf{A}^*$ can have at least one substantially non-zero value. At the same time, the number of elements with small non-zero values (those in J_2) can be large, but we control the absolute sum of their values so as to control the sparse estimation error of our estimators.

Theoretically, the value τ, defined in the sets J_1 and J_2, can be as large as $\min_{j\in J_1}|\xi_j^*|$. We can estimate τ as the smallest of $|\widehat{\xi}_j|$ for $j \in \widehat{J}_1$. We check this important sparsity assumption in practice, for the case when $M > 0$. If $\widehat{\mathbf{A}}$ is not very sparse (e.g., with non-zeros in every row), then we can split the data into a training and test set (e.g., with $T - T^{1/2}$ and $T^{1/2}$ data points respectively), and estimate the model again using the training set. We obtain estimators for the $\mathbf{y}_t$'s in the test set using the model trained from the training set. Out-sample estimation errors in the test set can then be obtained. If the level of this error is very different from the in-sample errors from the training set, then we suspect that the sparsity assumption may not be valid. We can do the same when $M = 0$, if $\widehat{\mathbf{A}}$ has many non-zero elements.

The independence of $\mathbf{B}_t$ and $\boldsymbol{\epsilon}_t$ in Assumption M3 ensures that $\mathbf{B}_t$ serves a function similar to an instrument for model (2.1). The tail condition in M3 implies that all the random variables involved have sub-exponential tails, so that exact normality is not required.

The assumption $\Theta^x_{m,2w} \le Cm^{-\alpha}$ in M4 essentially means that the strongest serial dependence for the x_{tj}'s with at least m time units apart is decaying polynomially, as m increases. Together with M3, they allow for the application of a Nagaev-type inequality in Lemma 1 in the supplemen-

tary material for our results to hold. Stationary Markov Chains and stationary linear processes are examples of time series that satisfy M4. See Chen et al. (2013).

3.3.2 Identification of the model

Assumption M5 is sufficient for the identification of model (3.3). Consider two sets of parameters $(\boldsymbol{\xi}^*, \boldsymbol{\delta}^*, \boldsymbol{\beta}^*)$ and $(\boldsymbol{\xi}^o, \boldsymbol{\delta}^o, \boldsymbol{\beta}^o)$ both satisfying model (3.3). Then

$$\mathbf{0} = \mathbf{B}^{\mathrm{T}}\mathbf{Z}(\boldsymbol{\xi}^* - \boldsymbol{\xi}^o) + \mathbf{B}^{\mathrm{T}}\mathbf{Z}\mathbf{V}_0(\boldsymbol{\delta}^* - \boldsymbol{\delta}^o) + \mathbf{B}^{\mathrm{T}}\mathbf{X}_{\boldsymbol{\beta}^* - \boldsymbol{\beta}^o}\mathrm{vec}(\mathbf{I}_N).$$

But

$$\begin{aligned}\mathbf{B}^{\mathrm{T}}\mathbf{X}_{\boldsymbol{\beta}^* - \boldsymbol{\beta}^o}\mathrm{vec}(\mathbf{I}_N) &= T^{-1/2}N^{-a/2}\begin{pmatrix} \sum_{t=1}^{T}(\mathbf{B}_t - \bar{\mathbf{B}})\boldsymbol{\gamma}\mathbf{x}_{t,1}^{\mathrm{T}}(\boldsymbol{\beta}^* - \boldsymbol{\beta}^o) \\ \vdots \\ \sum_{t=1}^{T}(\mathbf{B}_t - \bar{\mathbf{B}})\boldsymbol{\gamma}\mathbf{x}_{t,N}^{\mathrm{T}}(\boldsymbol{\beta}^* - \boldsymbol{\beta}^o) \end{pmatrix} \\ &= \mathbf{B}^{\mathrm{T}}\widetilde{\mathbf{X}}(\boldsymbol{\beta}^* - \boldsymbol{\beta}^o).\end{aligned}$$

Hence, with the definition of the set S in Assumption M5, we have

$$[\mathbf{B}^{\mathrm{T}}\mathbf{Z}_S \;\; \mathbf{B}^{\mathrm{T}}\mathbf{Z}\mathbf{V}_0 \;\; \mathbf{B}^{\mathrm{T}}\widetilde{\mathbf{X}}]\begin{pmatrix} \boldsymbol{\xi}_S^* - \boldsymbol{\xi}_S^o \\ \boldsymbol{\delta}^* - \boldsymbol{\delta}^o \\ \boldsymbol{\beta}^* - \boldsymbol{\beta}^o \end{pmatrix} = \mathbf{0}.$$

The above is true even for approximately sparse $\mathbf{A}^*$ and $\mathbf{A}^o$, since we assume that all the $o(1)$ elements (i.e., all those elements in J_2) are identified to 0, meaning that if both ξ_j^* and ξ_j^o are $o(1)$, then $\xi_j^* - \xi_j^o = 0$. Then the corresponding column in the matrix $\mathbf{B}^{\mathrm{T}}\mathbf{Z}$ can be removed. Taking expectation and multiplying $\mathbf{Q}^{\mathrm{T}}$ on both sides and then $(\mathbf{Q}^{\mathrm{T}}\mathbf{Q})^{-1}$, we arrive at $\boldsymbol{\xi}_S^* = \boldsymbol{\xi}_S^o$, $\boldsymbol{\delta}^* = \boldsymbol{\delta}^o$, and $\boldsymbol{\beta}^* = \boldsymbol{\beta}^o$.

It may seem that the assumption of being able to identify the "small" elements to 0 is strong under approximate sparsity. However, Theorem 3 does present a sign-consistent estimator which estimates all "small" elements to 0, with probability going to 1. Such an assumption is thus reasonable for our estimator.

Note that the matrix $\mathbf{Q}$ has size $N^2 \times (|S| + M + K)$. By Assumption M2 or M2' where there are only finite numbers of elements in each row of $\mathbf{A}^*$ and $\mathbf{A}^o$ belonging to J_1, we can see that $|S|$ is, at most, of order N. Hence assuming $\mathbf{Q}$ is of full rank is reasonable, since N^2 is then much larger than $|S| + M + K$, which is also of order N. In this sense, we see that the identification of the model parameters relies mainly on the sparsity of the sparse adjustment matrix.

3.3.3 Explanations for Assumptions R1 to R8, R5' and R8'

Assumption R1 essentially requires that each specification $\mathbf{W}_{0i}$ be different from each other, to a certain extent. This is intuitive, since if $\mathbf{W}_{0i}$ and $\mathbf{W}_{0j}$ are too similar to each other, the coefficients δ_i^* and δ_j^* are not well-defined. This will have a negative impact on the performance of our estimators. This assumption is analogous to requiring the columns in the design matrix to be linearly independent in a classical linear regression. As an example, if $\mathbf{W}_{0\ell}$ has an (i,j)th element being $d_{ij}^{-\ell}$ (excluding the diagonal which is 0) for $\ell = 1, \ldots, M$ with at least M of the d_{ij}'s different from each other, then each $\text{vec}(\mathbf{W}_{0i}^{\mathrm{T}})$ is linearly independent of the others. So $\mathbf{V}_0$, defined in (3.2), has full rank.

The assumptions on $\boldsymbol{\epsilon}_t$ and $\boldsymbol{\Sigma}_\epsilon$ in R2 are included mainly for convenience of the proof. The martingale difference assumption for $\boldsymbol{\epsilon}_t^*$ is a relaxation to independence. This assumption also allows the elements of $\boldsymbol{\epsilon}_t$ to have a general second-order moments structure, as long as $\big\|\boldsymbol{\Sigma}_\epsilon\big\|_\infty < \infty$. For instance, if $\boldsymbol{\epsilon}_t$ also follows a spatial lag model,

$$\boldsymbol{\epsilon}_t = \rho \mathbf{M} \boldsymbol{\epsilon}_t + \mathbf{u}_t,$$

where $\mathbf{u}_t$ has diagonal covariance matrix $\boldsymbol{\Sigma}_u$ with $\big\|\boldsymbol{\Sigma}_u\big\|_{\max} < C_u < \infty$, and $\mathbf{M}$ has $\big\|\mathbf{M}\big\|_\infty \leq 1$ with $|\rho| < 1$, then $\big\|\rho\mathbf{M}\big\|_\infty < 1$. Therefore,

$$\boldsymbol{\epsilon}_t = (\mathbf{I}_N - \rho\mathbf{M})^{-1}\mathbf{u}_t, \quad \text{with } \boldsymbol{\Sigma}_\epsilon = (\mathbf{I}_N - \rho\mathbf{M})^{-1}\boldsymbol{\Sigma}_u(\mathbf{I}_N - \rho\mathbf{M})^{-1}.$$

This spatial model for $\boldsymbol{\epsilon}_t$ is very common in spatial econometrics. It is clear that $\big\|\boldsymbol{\Sigma}_\epsilon\big\|_\infty \leq C_u(1 - \rho\big\|\mathbf{M}\big\|_\infty)^{-2} < \infty$.

Assumptions R3, R4, and R5 are closely related. They all paint a picture of how the exogenous variables in $\mathbf{B}_t$ are correlated with $\mathbf{X}_t$. Assumption R3 essentially says that covariance between a variable in $\mathbf{B}_t$ and one in $\mathbf{X}_t$ is finite uniformly as $N \to \infty$. Then for $1 \leq k \leq K$, considering the kth diagonal entry of $E(\mathbf{X}_t^{\mathrm{T}}\mathbf{B}_t)$ to be $\sum_{j=1}^N E(X_{t,jk}B_{t,jk})$ with each $E(X_{t,jk}B_{t,jk})$ finite, it is indeed reasonable to assume that each diagonal entry in the matrix is of order N, so that it is also reasonable to assume that this finite $K \times K$ matrix has all singular values of order N. This assumption is needed for the estimator $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta}(\widetilde{\boldsymbol{\theta}})$ to be well-defined in (2.7).

Assumption R4 is closely related to the identification condition M5, since the upper left block of $\mathbf{Q}^{\mathrm{T}}\mathbf{Q}$ is in fact $\mathbf{G}_{SS}$, and the identification condition already implies that $\mathbf{G}_{SS}$ has all eigenvalues uniformly bounded away from 0. Assumption R4 places more stringent condition, in the sense that the whole matrix $\mathbf{G}$ now has full rank. All eigenvalues of $\mathbf{G}$ are then uniformly bounded away from 0 and infinity. This is because it can avoid the need for a restricted eigenvalue condition which is unnecessarily complicated in our context.

Assumption R5 essentially describes how each row of variables in $\mathbf{B}_t$ are correlated with different rows of variables in $\mathbf{X}_t$. The rate N^a is closely related to Assumption R4. This can be

seen by noting that each block $E(T^{-1}\sum_{t=1}^{T}(\mathbf{B}_t - \bar{\mathbf{B}})\boldsymbol{\gamma}\mathbf{y}_t^{\mathrm{T}})$ in $\mathbf{G}$ is asymptotically the same as $(I_N \otimes \boldsymbol{\gamma}^{\mathrm{T}})\mathbf{L}_0(\mathbf{I}_N \otimes \boldsymbol{\beta}^*)\boldsymbol{\Pi}^{*\mathrm{T}}$, where $\mathbf{L}_0 = E(\mathrm{vec}(\mathbf{B}_t^{\mathrm{T}})\mathrm{vec}(\mathbf{X}_t^{\mathrm{T}})^{\mathrm{T}})$. Note that the (i,j)th $K \times K$ block in $\mathbf{L}_0$ is indeed $E(\mathbf{b}_{t,i}\mathbf{x}_{t,j}^{\mathrm{T}})$. Assumption R4 then essentially says that both

$$\max_{1\leq i\leq N}\sum_{j=1}^{N}\left\|E(\mathbf{b}_{t,i}\mathbf{x}_{t,j}^{\mathrm{T}})\right\|_{\max}^2, \quad \max_{1\leq j\leq N}\sum_{i=1}^{N}\left\|E(\mathbf{b}_{t,i}\mathbf{x}_{t,j}^{\mathrm{T}})\right\|_{\max}^2$$

are of order similar to N^a. Assumption R5 can then be seen as the regularity condition such that the upper bounds (with square removed for ease of proofs) are exactly of order N^a. With this, we can derive that $\left\|E(\mathbf{X}_t \otimes \mathbf{B}_t\boldsymbol{\gamma})\right\|_1$ has order at most N^{1+a}. This particular condition is needed for proving the rates of $\left\|\widetilde{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\right\|_1$ in Theorem 1.

When $M = 0$, this term is not required, and instead we have $n = O(N)$ in Assumption M2'. We then need the rate in Assumption R5' to guarantee correct asymptotic normality formulae of the adaptive LASSO estimators $\widehat{\boldsymbol{\xi}}$ and $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}(\widehat{\boldsymbol{\xi}})$. This rate does not contradict the two rates listed in Assumption R5 and R5', but refines them. The two rates in R5 and R5' mean that some rows of $\mathbf{B}_t$ are strongly correlated with a certain of rows of $\mathbf{X}_t$. Then the rate $\left\|E(\mathbf{B}_t \otimes \boldsymbol{\Pi}^*\mathbf{X}_t\boldsymbol{\beta}^*)\right\|_1 = O(N)$ says that the number of such rows in $\mathbf{B}_t$ is finite, and the majority of the rows of $\mathbf{B}_t$ are not always strongly correlated with many rows of $\mathbf{X}_t$. For instance, there can be some universal economic variables in some rows of $\mathbf{B}_t$, but the majority of the variables are local, so they do not correlate strongly with too many rows of $\mathbf{X}_t$. We can check this assumption using the estimated spatial weight matrix and estimated regression parameters.

Assumption R6 gives a rate for the singular values of $\mathrm{var}(\mathbf{B}_{t,k})$. This is important in certain asymptotic normality results. The rate N^b, possibly different from N^a, is reasonable as well, since the way that $\mathbf{B}_t$ and $\mathbf{X}_t$ are correlated does not directly indicate how the variables in $\mathbf{B}_t$ itself are correlated. That is, unless when $\mathbf{B}_t = \mathbf{X}_t$ when $\mathbf{X}_t$ itself is exogenous, in which case $b = a$. The variance-covariance matrix dominating the lag-τ auto-covariances enables easier presentation of rates of convergence in asymptotic normality.

Assumption R7 spells out the rate of the penalization parameters for all adaptive LASSO estimators in this paper to be (partial) sign consistent, with non-zeros enjoying asymptotic normality. This also makes it easier to grid search for the best tuning parameter using the BIC criterion (4.17), to be introduced in Section 4.

For Assumption R8, when $M > 0$, the first line of rates are needed for the oracle inequality and preliminary rates of convergence for $\left\|\widetilde{\boldsymbol{\xi}}_{J_1} - \boldsymbol{\xi}_{J_1}^*\right\|$. This inequality is used for the rest of the proofs of all other theorems. Since Assumption M2 has $n = o(N^{1/2})$, $c_T n = o(1)$ potentially still allows for N to grow faster than T, while $nN^{a-1} = o(1)$ means $a \leq 1/2$. If $M = 0$, we have $n = O(N)$ from Assumption M2', and so $c_T n = o(1)$ means N has to grow slower than $T^{1/2}$. This is not surprising, as potentially we have a large number of non-zeros to estimate in $\mathbf{A}^*$. Since we are not making any structural assumptions on $\mathbf{A}^*$ other than sparseness, having more non-zeros

estimated in $\mathbf{A}^*$ means N cannot grow too fast relative to T. Also in R8', the rate related to nN^{a-1} is dropped, since when $M > 0$ this rate controls the rate of convergence of $\widetilde{\boldsymbol{\delta}}$, which is identically zero when $M = 0$. The addition of $N^{a-b}\log(T \vee N) = o(1)$ is needed for the correct formula for the asymptotic normality of $\widehat{\boldsymbol{\beta}}$.

The second line of rates in Assumption R8 (and R8') are required for controlling the dominating terms in all asymptotic normality results, while the other two rates involving $\boldsymbol{\xi}^*_{J_2}$ provide partial sign consistency of $\widehat{\boldsymbol{\xi}}$. If $\mathbf{A}^*$ is exactly sparse, then $J_2 = \phi$ and $\boldsymbol{\xi}^*_{J_2} = \mathbf{0}$, so that the rates are trivially satisfied. For instance, if $M > 0$, $a = b = 1/2$, and $\mathbf{A}^*$ is exactly sparse, then N can grow faster than T depending on the number of large non-zero elements in $\mathbf{A}^*$.

3.3.4 Extension: allowing past $\mathbf{y}_t$

The assumptions above allow $\mathbf{X}_t$ to contain past values of $\mathbf{y}_t$. If $\mathbf{X}_t = (\mathbf{y}_{t-1}, \ldots, \mathbf{y}_{t-d}, \mathbf{z}_t)$ for instance, where $\mathbf{z}_t$ contains covariates other than $\{\mathbf{y}_t\}$, then

$$\mathbf{X}_t\boldsymbol{\beta}^* = \sum_{j=1}^{d} \beta_j^* \mathbf{y}_{t-j} + \mathbf{z}_t\boldsymbol{\beta}_2^*,$$

with $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_d^*, \boldsymbol{\beta}_2^{*\mathrm{T}})^{\mathrm{T}}$. Hence model (2.1) becomes

$$\mathbf{y}_t = \boldsymbol{\mu}^* + \mathbf{W}^*\mathbf{y}_t + \sum_{j=1}^{d} \beta_j^* \mathbf{y}_{t-j} + \mathbf{z}_t\boldsymbol{\beta}_2^* + \boldsymbol{\epsilon}_t.$$

The reduced form model in (3.14) then becomes

$$\mathbf{y}_t = \boldsymbol{\Pi}^*\boldsymbol{\mu}^* + \sum_{j=1}^{d} \beta_j^* \boldsymbol{\Pi}^* \mathbf{y}_{t-j} + \boldsymbol{\Pi}^*(\mathbf{z}_t\boldsymbol{\beta}_2^* + \boldsymbol{\epsilon}_t). \tag{3.15}$$

This model has a vector autoregressive part with coefficient matrices $\beta_j^*\boldsymbol{\Pi}^*$. It is similar to vector autoregressions (VARs) requiring the estimation of dN^2 parameters. In typical macroeconomic applications, the number of time periods T might be small relative to the number of parameters. In particular, Kock and Callot (2015) demonstrate Oracle properties of LASSO and Adaptive LASSO estimators in the context of high-dimensional vector autoregresions. See also Medeiros and Mendes (2016). Model (3.15) is more general, since it allows for lagged $\mathbf{y}_t$ along with covariates $\mathbf{X}_t$ and instrumental variables. It also complements recent research on global VARs. Pesaran et al. (2004) consider country-specific VARs, interlinked by cross-country spillovers and international fluctuations in the global economy. Other approaches include the Bayesian modelling of Bańbura et al. (2010) and data-reduction techniques, such as the factor-augmented VAR of Bernanke et al. (2005).

3.4 Main results

We first present the rates related to our estimators for an arbitrary sparse adjustment matrix.

Theorem 1 *Let all the assumptions in Section 3.3 hold (excluding M2', R5', R8'). Moreover, let* $\alpha > 1/2 - 1/w$ *in Assumption M4, and* $N = o(T^{w/4-1/2}\log^{w/4}(T))$. *Let* $\widetilde{\mathbf{A}}$ *be any estimator of the sparse adjustment* $\mathbf{A}^*$ *(not necessarily a LASSO estimator). Then the estimator* $\widetilde{\boldsymbol{\delta}}$ *in (3.5) and* $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta}(\widetilde{\boldsymbol{\theta}})$ *in (3.4), with* $\widetilde{\boldsymbol{\theta}} = (\widetilde{\boldsymbol{\xi}}^{\mathrm{T}}, \widetilde{\boldsymbol{\delta}}^{\mathrm{T}})^{\mathrm{T}}$ *and* c_T *defined in Assumption R7, satisfy*

$$\left\|\widetilde{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\right\|_1 = O_p(c_T N^{-\frac{1}{2}+\frac{1}{2w}} + N^{-1}\left\|\widetilde{\boldsymbol{\xi}} - \boldsymbol{\xi}^*\right\|_1) = \left\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_1.$$

If $M = 0$, *the above still hold if Assumptions M2, R5, and R8 are replaced by M2', R5', and R8', respectively.*

The quality of both $\widetilde{\boldsymbol{\delta}}$ and $\widetilde{\boldsymbol{\beta}}$ is dependent on $\widetilde{\boldsymbol{\xi}}$, and hence $\widetilde{\mathbf{A}}$, as shown by the above bound. If our initial specifications $\mathbf{W}_{01}, \ldots, \mathbf{W}_{0M}$ are insufficient to form a good linear combination for estimating the spatial weight matrix, then we need more non-zero adjustments on $\mathbf{A}^*$. Estimating too many of them, however, inflates the error bound $\left\|\widetilde{\boldsymbol{\xi}} - \boldsymbol{\xi}^*\right\|_1$. Also, the constant w can be chosen to be large enough to satisfy $N = o(T^{w/4-1/2}\log^{w/4}(T))$, so that $N^{1/(2w)}$ can grow very slowly compared to $N^{1/2}$.

Theorem 2 *Let the assumptions in Section 3.3 and in Theorem 1 hold. Then the LASSO solution* $\widetilde{\boldsymbol{\xi}}$ *satisfies*

$$\begin{aligned}\left\|\widetilde{\boldsymbol{\xi}} - \boldsymbol{\xi}^*\right\|_1 &= O_p(c_T N^{\frac{1}{2}+\frac{1}{2w}} + n^{1/2}\left\|\widetilde{\boldsymbol{\xi}}_{J_1} - \boldsymbol{\xi}^*_{J_1}\right\|),\\ \left\|\widetilde{\boldsymbol{\xi}}_{J_1} - \boldsymbol{\xi}^*_{J_1}\right\| &= O_p(c_T n^{1/2}(1 + N^{a-\frac{1}{2}+\frac{1}{2w}}) + c_T N^{\frac{1}{2w}}(N^{a/2} + c_T^{1/2}N^{1/2})).\end{aligned}$$

For the LASSO estimators $\widetilde{\boldsymbol{\beta}}$ *and* $\widetilde{\boldsymbol{\delta}}$,

$$\left\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_1 = O_p(c_T N^{-\frac{1}{2}+\frac{1}{2w}}) = \left\|\widetilde{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\right\|_1.$$

If $M = 0$ *and Assumptions M2, M5, and M8, are replaced by M2', R5', and R8' respectively, the rate for* $\left\|\widetilde{\boldsymbol{\xi}} - \boldsymbol{\xi}^*\right\|_1$ *is the same as above, but with*

$$\left\|\widetilde{\boldsymbol{\xi}}_{J_1} - \boldsymbol{\xi}^*_{J_1}\right\| = O_p(c_T n^{1/2}(1 + N^{-\frac{1}{2}+\frac{1}{2w}}) + c_T^{3/2}N^{\frac{1}{2}+\frac{1}{2w}}), \quad \left\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_1 = O_p(c_T).$$

Although $\left\|\widetilde{\boldsymbol{\xi}} - \boldsymbol{\xi}^*\right\|_1$ may not necessarily converge to 0, the L_2 error $\left\|\widetilde{\boldsymbol{\xi}}_{J_1} - \boldsymbol{\xi}^*_{J_1}\right\|$ does indeed go to 0 in probability. Our empirical results confirm this. The term $\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ has a slower rate of convergence when $M = 0$, since from Assumption M2', more non-zero elements are in $\mathbf{A}^*$ (of order N) to be estimated. These bounds are important stepping stones for proving Theorems 3, 4, and 5 for all adaptive LASSO estimators.

Theorem 3 *(Oracle Property for $\widehat{\boldsymbol{\xi}}$) Let the assumptions in Section 3.3 and in Theorem 1 hold (excluding M2', R5', and R8'). Then, with large enough T and N, having probability going to 1, $\widehat{\boldsymbol{\xi}}$ in (3.6) satisfies*

$$\operatorname{sign}(\widehat{\boldsymbol{\xi}}_{J_1}) = \operatorname{sign}(\boldsymbol{\xi}^*_{J_1}), \quad \widehat{\boldsymbol{\xi}}_{J_0 \cup J_2} = \mathbf{0},$$

where J_0, J_1, and J_2 are defined in Assumption M2. Moreover, define the predictive dependence measures

$$P_0^b(B_{t,sk}) = E(B_{t,sk}|\mathcal{G}_0) - E(B_{t,sk}|\mathcal{G}_{-1}), \;\; P_0^\epsilon(\epsilon_{tj}) = E(\epsilon_{tj}|\mathcal{H}_0) - E(\epsilon_{tj}|\mathcal{H}_{-1}),$$

where $\mathcal{G}_t$ and $\mathcal{H}_t$ are defined just after Equation (3.11). Assume

$$\sum_{t\geq 0} \max_{1\leq k\leq K} \max_{1\leq s\leq N} \left\|P_0^b(B_{t,sk})\right\| < \infty, \quad \sum_{t\geq 0} \max_{1\leq j\leq N} \left\|P_0^\epsilon(\epsilon_{tj})\right\| < \infty. \tag{3.16}$$

Then for $\mathbf{M} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_m)^{\mathrm{T}}$ with m finite and $\left\|\boldsymbol{\alpha}_i\right\|_1 < \infty$, we have

$$T^{1/2}(\mathbf{MR\Sigma R}^{\mathrm{T}}\mathbf{M}^{\mathrm{T}})^{-1/2}\mathbf{M}(\widehat{\boldsymbol{\xi}}_{J_1} - \boldsymbol{\xi}^*_{J_1} + \mathbf{G}^{-1}_{J_1J_1}\lambda_T\mathbf{g}_{J_1}) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \mathbf{I}_m),$$

where $\mathbf{R} = N^{-a}\mathbf{G}^{-1}_{J_1J_1}E(T^{-1}\mathbf{Z}^{\mathrm{T}}_{J_1}(\mathbf{B}_{\boldsymbol{\gamma}} - \bar{\mathbf{B}}_{\boldsymbol{\gamma}}))$, and $\mathbf{\Sigma} = \sum_\tau E(\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}^{\mathrm{T}}_{t+\tau}) \otimes E((\mathbf{B}_t - \boldsymbol{\mu}_b)\boldsymbol{\gamma}\boldsymbol{\gamma}^{\mathrm{T}}(\mathbf{B}_{t+\tau} - \boldsymbol{\mu}_b)^{\mathrm{T}})$. The vector $\mathbf{g}_{J_1}$ is a vector of 1 or -1, depending on whether the corresponding element in $\boldsymbol{\xi}^*_{J_1}$ is positive or negative.

If $M = 0$, the above still holds for $\widehat{\boldsymbol{\xi}}$ in (3.9) if Assumptions M2, R5, and R8 are replaced by M2', R5', and R8', respectively.

The predictive dependence measure, introduced in Definition 2 of Wu (2011), quantifies the degree of dependence of outputs on inputs in physical systems, similar to our variables defined in (3.11). To draw inference on the non-zero sparse adjustments on $\mathbf{A}^*$, we can use the sign consistency of $\widehat{\boldsymbol{\xi}}$ and the asymptotic normality of $\widetilde{\boldsymbol{\xi}}_{J_1}$. We can estimate the matrices $\mathbf{R}$ and $\mathbf{\Sigma}$ using the corresponding sample autocovariance estimators. Since $\{\boldsymbol{\epsilon}_t\}$ is unobserved, we replace this by $\{\widehat{\boldsymbol{\epsilon}}_t\}$ obtained as the residuals of model (2.1) with $\boldsymbol{\mu}^*$, $\mathbf{A}^*$, $\boldsymbol{\delta}^*$, and $\boldsymbol{\beta}^*$ replaced by $\widehat{\boldsymbol{\mu}}$, $\widehat{\mathbf{A}}$, $\widehat{\boldsymbol{\delta}}$, and $\widehat{\boldsymbol{\beta}}$, respectively. Note that $\mathbf{R}$ is in fact independent of the unknown index a. Since $\mathbf{G}$ also has the term N^{-a}, it cancels out the N^{-a} in the definition of $\mathbf{R}$. The term $\boldsymbol{\xi}^*_{J_1}$ is a vector of non-zero constants while on $\mathcal{M}$, $\left\|\mathbf{G}^{-1}_{J_1J_1}\lambda_T\mathbf{g}_{J_1}\right\|_{\max} = O(c_T) = o(1)$ (see the proof of the Theorem in the supplementary materials for more details), we can ignore the term $\mathbf{G}^{-1}_{J_1J_1}\lambda_T\mathbf{g}_{J_1}$ in using the asymptotic normality result in construction of confidence intervals.

To illustrate, if $(\widehat{\mathbf{A}})_{23}$ (corresponding to $\widehat{\xi}_{N+3}$) and $(\widehat{\mathbf{A}})_{35}$ (corresponding to $\widehat{\xi}_{2N+5}$) are non-zero and we want to make inference on them, then we can set $m = 2$ and $\mathbf{M} = (\mathbf{e}_{N+3}, \mathbf{e}_{2N+5})^{\mathrm{T}}$ where $\mathbf{e}_i \in \mathbb{R}^{N^2}$ is 0 everywhere except 1 at the ith position. The asymptotic bivariate normality of the two estimators can then be established, and a confidence region can be constructed.

In practice, we can find a reasonable cut-off on τ in calculating $\mathbf{\Sigma}$. See Remark 2 in Section 4

for more practical details on this. The rate of convergence of each $\widehat{\boldsymbol{\xi}}_j$ for $j \in J_1$ can be deduced to be $T^{-1/2}N^{-(a-b)/2}$ from the asymptotic normality result. This is done by using Assumption R6 to deduce that $N^{-b}\boldsymbol{\Sigma}$ has uniformly bounded eigenvalues from 0 and infinity.

The asymptotic normality for $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}(\widehat{\boldsymbol{\xi}}, \widetilde{\boldsymbol{\delta}})$ (when $M > 0$) and $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}(\widehat{\boldsymbol{\xi}})$ (when $M = 0$) appear in the following theorem.

Theorem 4 *Let the assumptions in Section 3.3 and in Theorem 1 hold (excluding M2', R5', and R8'). Assume that the predictive dependence measures $P_0^b(B_{tk})$ and $P_0^\epsilon(\epsilon_{tj})$ are as defined in Theorem 3 with the same assumptions (3.16) applied. Then for $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}(\widehat{\boldsymbol{\xi}}, \widetilde{\boldsymbol{\delta}})$, we have*

$$T^{1/2}\mathbf{S}_0^{-1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \mathbf{I}_K),$$

where, defining $\mathbf{R}_0 = (E(\mathbf{X}_t^{\mathrm{T}}\mathbf{B}_t)E(\mathbf{B}_t^{\mathrm{T}}\mathbf{X}_t))^{-1}E(\mathbf{X}_t^{\mathrm{T}}\mathbf{B}_t)$,

$$\mathbf{S}_0 = \sum_\tau \mathbf{R}_0 E(\mathbf{B}_t^{\mathrm{T}}\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_{t+\tau}^{\mathrm{T}}\mathbf{B}_{t+\tau})\mathbf{R}_0^{\mathrm{T}}.$$

If $M = 0$ and Assumptions M2, R5, and R8 are replaced by M2', R5', and R8', respectively, then for $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}(\widehat{\boldsymbol{\xi}})$,

$$T^{1/2}(\mathbf{K}_0\mathbf{R}\boldsymbol{\Sigma}\mathbf{R}^{\mathrm{T}}\mathbf{K}_0^{\mathrm{T}})^{-1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \mathbf{I}_K),$$

where $\mathbf{R}$ and $\boldsymbol{\Sigma}$ are defined in Theorem 3, and

$$\mathbf{K}_0 = [E(\mathbf{X}_t^{\mathrm{T}}\mathbf{B}_t)E(\mathbf{B}_t^{\mathrm{T}}\mathbf{X}_t)]^{-1}E(\mathbf{X}_t^{\mathrm{T}}\mathbf{B}_t)E(\mathbf{B}_t^{\mathrm{T}} \otimes \boldsymbol{\beta}^{*\mathrm{T}}\mathbf{X}_t^{\mathrm{T}}\boldsymbol{\Pi}^{*\mathrm{T}}).$$

It is unsurprising that the asymptotic normality results for $M > 0$ and $M = 0$ are different. When $M > 0$ we have $n = o(N^{1/2})$, but it is $n = O(N)$ when $M = 0$. The dominating terms under these two scenarios are different, and so the asymptotic covariance matrices, as well as the rates of convergence, are also different. As discussed under Theorem 3, we can estimate $\mathbf{S}_0$, $\mathbf{R}$, $\boldsymbol{\Sigma}$, and $\mathbf{K}_0$ using appropriate sample autocovariance matrices, with $\boldsymbol{\epsilon}_t$, $\boldsymbol{\beta}^*$, and $\boldsymbol{\Pi}^*$ replaced by $\widehat{\boldsymbol{\epsilon}}_t$, $\widehat{\boldsymbol{\beta}}$, and $(\mathbf{I}_N - \widehat{\mathbf{W}})^{-1}$, respectively. Similar to estimating $\boldsymbol{\Sigma}$, we can find a reasonable cut-off for the sum in the definition of $\mathbf{S}_0$. See Remark 2 in Section 4 for more practical details on this.

Next, we present the oracle property of $\widehat{\boldsymbol{\delta}}$ defined in (3.7) in the following theorem.

Theorem 5 *(Oracle property for $\widehat{\boldsymbol{\delta}}$) Let the assumptions in Section 3.3 and in Theorem 1 hold (excluding M2', R5', and R8'). We then have, as $T, N \to \infty$, with probability approaching 1,*

$$\mathrm{sign}(\widehat{\boldsymbol{\delta}}_H) = \mathrm{sign}(\boldsymbol{\delta}_H^*), \quad \widehat{\boldsymbol{\delta}}_{H^c} = \mathbf{0}, \quad \textit{where } H = \{j : \delta_j^* \neq 0\}.$$

Moreover, with the predictive dependence measures $P_0^b(B_{tk})$ and $P_0^\epsilon(\epsilon_{tj})$ as defined in Theorem 3

with the same assumptions (3.16) applied, we have

$$T^{1/2}(\mathbf{R}_1\mathbf{S}_{\boldsymbol{\gamma}}\mathbf{S}_0\mathbf{S}_{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{R}_1^{\mathrm{T}})^{-1/2}(\widehat{\boldsymbol{\delta}}_H - \boldsymbol{\delta}_H^*) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \mathbf{I}_{|H|}),$$

where $\mathbf{R}_1 = [(\mathbf{H}_{10} - \mathbf{H}_{20})_H^{\mathrm{T}}(\mathbf{H}_{10} - \mathbf{H}_{20})_H]^{-1}(\mathbf{H}_{10} - \mathbf{H}_{20})_H^{\mathrm{T}}$, *with* $\mathbf{H}_{10}, \mathbf{H}_{20}$ *and* $\mathbf{S}_{\boldsymbol{\gamma}}$ *defined as*

$$\begin{aligned}
\mathbf{H}_{10} &= \Big(\mathbf{I}_N \otimes (\mathbf{I}_N \otimes \boldsymbol{\gamma}^{\mathrm{T}})E(vec(\mathbf{B}_t^{\mathrm{T}})vec(\mathbf{X}_t^{\mathrm{T}})^{\mathrm{T}})(\mathbf{I}_N \otimes \boldsymbol{\beta}^*)\boldsymbol{\Pi}^{*\mathrm{T}}\Big)\mathbf{V}_0, \\
\mathbf{H}_{20} &= E(\mathbf{X}_t \otimes \mathbf{B}_t\boldsymbol{\gamma})\Big(E(\mathbf{X}_t^{\mathrm{T}}\mathbf{B}_t)E(\mathbf{B}_t^{\mathrm{T}}\mathbf{X}_t)\Big)^{-1}E(\mathbf{X}_t^{\mathrm{T}}\mathbf{B}_t)\Big(\mathbf{V}_{\mathbf{W}_{01}^{\mathrm{T}}}^{\mathrm{T}} \cdots \mathbf{V}_{\mathbf{W}_{0M}^{\mathrm{T}}}^{\mathrm{T}}\Big)(\mathbf{I}_M \otimes \mathbf{U}_0\mathbf{V}_{\boldsymbol{\Pi}^*}\boldsymbol{\beta}^*), \text{ and} \\
\mathbf{S}_{\boldsymbol{\gamma}} &= \Big(cov(\mathbf{x}_{t,1}, \mathbf{b}_{t,1})\boldsymbol{\gamma}, \ldots, cov(\mathbf{x}_{t,1}, \mathbf{b}_{t,N})\boldsymbol{\gamma}, \ldots, cov(\mathbf{x}_{t,N}, \mathbf{b}_{t,1})\boldsymbol{\gamma}, \ldots, cov(\mathbf{x}_{t,N}, \mathbf{b}_{t,N})\boldsymbol{\gamma}\Big)^{\mathrm{T}}.
\end{aligned}$$

The definition of $\mathbf{S}_0$ *is as in Theorem 4, while* $\mathbf{U}_0 = \mathbf{I}_N \otimes E(\mathbf{b}_t\mathbf{x}_t^{\mathrm{T}})$ *and*

$$\mathbf{V}_{\boldsymbol{\Pi}^*} = \begin{pmatrix} \mathbf{I}_K \otimes \boldsymbol{\pi}_1^* \\ \vdots \\ \mathbf{I}_K \otimes \boldsymbol{\pi}_N^* \end{pmatrix},$$

where $\boldsymbol{\pi}_j^{*\mathrm{T}}$ *is the* j*th row of* $\boldsymbol{\Pi}^*$. *We also assume that the smallest eigenvalue of* $\mathbf{R}_1\mathbf{S}_{\boldsymbol{\gamma}}\mathbf{S}_{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{R}_1^{\mathrm{T}}$ *is of constant order. In particular, we have*

$$T^{1/2}s_1^{-1/2}(\widehat{\rho} - \rho^*) = T^{1/2}s_1^{-1/2}\mathbf{1}_{|H|}^{\mathrm{T}}(\widehat{\boldsymbol{\delta}}_H - \boldsymbol{\delta}_H^*) \xrightarrow{\mathcal{D}} N(0, 1),$$

where $s_1 = \mathbf{1}_{|H|}^{\mathrm{T}}\mathbf{R}_1\mathbf{S}_{\boldsymbol{\gamma}}\mathbf{S}_0\mathbf{S}_{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{R}_1^{\mathrm{T}}\mathbf{1}_{|H|}$.

The above theorem is important in determining which potential specifications of spatial weight matrices are important and which are not, and how important each specification is. Similar to $\widehat{\boldsymbol{\xi}}_{J_1}$, hypothesis testing can be conducted and confidence regions constructed for the elements in $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\delta}}$. Various autocovariance matrices are estimated from the data and their respective asymptotic normality is used.

It may seem that expecting the smallest eigenvalue of $\mathbf{R}_1\mathbf{S}_{\boldsymbol{\gamma}}\mathbf{S}_{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{R}_1^{\mathrm{T}}$ to have a constant order is a strong assumption. However, in the proof of Theorem 5 in the supplementary material, we show that the largest eigenvalue of $\mathbf{R}_1\mathbf{S}_{\boldsymbol{\gamma}}\mathbf{S}_{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{R}_1^{\mathrm{T}}$ has a constant order. Considering the matrix is $K \times K$ and has a constant size, this assumption is not particularly strong.

Theorem 6 *Let the assumptions in Section 3.3 and in Theorem 1 hold. Then*

$$\big\|\widehat{\mathbf{W}} - \mathbf{W}^*\big\| = O_p(c_T) = \big\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}^*\big\|_{\max},$$

where $\widehat{\mathbf{W}} = \widehat{\mathbf{A}} + \sum_{r=1}^{M}\widehat{\delta}_r\mathbf{W}_{0r}$, $\widehat{\boldsymbol{\mu}} = (\mathbf{I}_N - \widehat{\mathbf{W}})\bar{\mathbf{y}} - \bar{\mathbf{X}}\widehat{\boldsymbol{\beta}}$. *The rate for the spatial fixed effect,* $\big\|\widehat{\boldsymbol{\Pi}}\widehat{\boldsymbol{\mu}} - \boldsymbol{\Pi}^*\boldsymbol{\mu}^*\big\|_{\max}$, *is the same.*

The above gives a rate of convergence for the spectral norm of the error for the estimated spatial weight matrix, which can be of independent interest.

4 Practical Implementation

In this section we provide details of the block-coordinate descent (BCD) algorithm used to conduct the minimization of (3.4)-(3.8). The optimization problem with respect to $(\boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\delta})$ is not convex. This gives rise to significant computational challenges. When $M = 0$, solving (3.10) does not involve any steps in the BCD algorithm, but just like step 3' (see below) we need to utilize the LASSO solution from the LASSO stage to solve for the adaptive LASSO problem. Hence the details for solving for $\widehat{\boldsymbol{\xi}}$ in (3.10) when $M = 0$ is omitted.

We make use of the fact that when given any two of the three variables, the problem is convex in the remaining variable. For example, given $(\boldsymbol{\xi}, \boldsymbol{\beta})$, the optimization problem is convex in $\boldsymbol{\delta}$. While it is difficult to establish global convergence of the BCD algorithm without convexity, each iteration delivers an improvement of the objective functions since given one parameter, the objective functions are convex in the others. The BCD algorithm is closely related to the Iterative Coordinate Descent of Fan and Lv (2011), and is also discussed by Friedman et al. (2010) and Dicker et al. (2013).

The Block Coordinate Descent Algorithm (LASSO stage)

0. Set an initial value $\boldsymbol{\xi}^{(0)}$, for example $\boldsymbol{\xi}^{(0)} = 0$.

1. At iteration r, update $\boldsymbol{\beta}^{(r)}$ according to the closed-form expression

$$\boldsymbol{\beta}^{(r)} = (\mathbf{X}^{\mathrm{T}}\mathbf{B}^{v}\mathbf{B}^{v\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{B}^{v}\mathbf{B}^{v\mathrm{T}}\left(\mathbf{I}_{TN} - (\mathbf{I}_T \otimes \mathbf{A}^{(r)}) - \sum_{i=1}^{M}\delta_i^{(r-1)}\mathbf{W}_{0i}^{\otimes}\right)\mathbf{y}^{v}.$$

2. Update $\boldsymbol{\delta}^{(r)}$ as a function of $\boldsymbol{\beta}^{(r)}$ and $\boldsymbol{\xi}^{(r)}$, with

$$\begin{aligned}\boldsymbol{\delta}^{(r)} = {} & \Big[(\mathbf{H} - \mathbf{B}^{\mathrm{T}}\mathbf{Z}\mathbf{V}_0)^{\mathrm{T}}(\mathbf{H} - \mathbf{B}^{\mathrm{T}}\mathbf{Z}\mathbf{V}_0)\Big]^{-1}(\mathbf{H} - \mathbf{B}^{\mathrm{T}}\mathbf{Z}\mathbf{V}_0)^{\mathrm{T}} \\ & \cdot(\mathbf{B}^{\mathrm{T}}\mathbf{Z}\boldsymbol{\xi}^{(r)} - \mathbf{B}^{\mathrm{T}}\mathbf{y} + \mathbf{K}(\mathbf{I}_{TN} - (\mathbf{I}_T \otimes \mathbf{A}^{(r)}))\mathbf{y}^{v}).\end{aligned}$$

3. Using the Least Angle Regression (LARS), solve sequentially for each row of $\mathbf{A}$, denoted η_j^{T}, fixing all the remaining $(1, \ldots, j-1, j+1, \ldots, n)$ rows, for $j = 1, \ldots, N$. That is, solve

$$\eta_j^{(r)} = \arg\min_{\eta_j} \frac{1}{2T}\left\|\mathbf{B}^{\mathrm{T}}\mathbf{y} - \mathbf{B}^{\mathrm{T}}\mathbf{Z}\eta - \mathbf{B}^{\mathrm{T}}\mathbf{Z}\mathbf{V}_0\boldsymbol{\delta}^{(r-1)} - \mathbf{B}^{\mathrm{T}}\mathbf{X}_{\boldsymbol{\beta}^{(r-1)}}\mathrm{vec}(\mathbf{I}_N)\right\|^2 + \lambda_T\left\|\eta\right\|_1,$$

where $\eta = (\eta_1^{(r-1)\mathrm{T}}, \ldots, \eta_{j-1}^{(r-1)\mathrm{T}}, \eta_j, \eta_{j+1}^{(r-1)\mathrm{T}}, \ldots, \eta_n^{(r-1)\mathrm{T}})^{\mathrm{T}}$, subject to the constraints as stated in (2.7). We obtain $\boldsymbol{\xi}^{(r)} = (\eta_1^{(r)\mathrm{T}}, \ldots, \eta_n^{(r)\mathrm{T}})^{\mathrm{T}}$.

4. Repeat steps 1-3, until $\left\|\xi^{(r)}-\xi^{(r-1)}\right\|$ is smaller than some predetermined number. The LASSO solution is $\widetilde{\boldsymbol{\xi}}=\boldsymbol{\xi}^{(r)}$, $\widetilde{\boldsymbol{\beta}}=\boldsymbol{\beta}^{(r)}$ and $\widetilde{\boldsymbol{\delta}}=\boldsymbol{\delta}^{(r)}$. We can construct $\widetilde{\mathbf{A}}$ from $\widetilde{\boldsymbol{\xi}}=\text{vec}(\widetilde{\mathbf{A}}^{\text{T}})$.

With the LASSO solution, compute the final adaptive LASSO stage, which is obtained by setting the initial values as $\widetilde{\boldsymbol{\xi}}$, $\widetilde{\boldsymbol{\beta}}$, and $\widetilde{\boldsymbol{\delta}}$, and repeating the steps above with the following modifications.

The Block Coordinate Descent Algorithm (Adaptive LASSO stage)

2′. Introduce penalization for δ:

$$\boldsymbol{\delta}^{(r)} = \arg\min_{\boldsymbol{\delta}}\frac{1}{2T}\left\|\mathbf{B}^{\text{T}}\mathbf{y}-\mathbf{B}^{\text{T}}\mathbf{Z}\boldsymbol{\xi}^{(r-1)}-\mathbf{g}^{(r-1)}-(\mathbf{B}^{\text{T}}\mathbf{Z}\mathbf{V}_0-\mathbf{H})\boldsymbol{\delta}\right\|^2+\lambda_T'\mathbf{u}^{\text{T}}|\boldsymbol{\delta}|,$$

where $\mathbf{u}=(|\widetilde{\delta}_1|^{-1},\ldots,|\widetilde{\delta}_M|^{-1})^{\text{T}}$, $\mathbf{g}^{(r-1)}$ is defined similar to $\widehat{\mathbf{h}}$ in (3.8) with $\widehat{\mathbf{A}}$ replaced by $\mathbf{A}^{(r-1)}$, and the above is subjected to the constraints as stated in (3.8).

3′. The adaptive LASSO objective function observes penalty $\lambda_T\mathbf{v}^{\text{T}}|\eta|$, that is,

$$\eta_j^{(r)} = \arg\min_{\eta_j}\frac{1}{2T}\left\|\mathbf{B}^{\text{T}}\mathbf{y}-\mathbf{B}^{\text{T}}\mathbf{Z}\eta-\mathbf{B}^{\text{T}}\mathbf{Z}\mathbf{V}_0\boldsymbol{\delta}^{(r-1)}-\mathbf{B}^{\text{T}}\mathbf{X}_{\boldsymbol{\beta}^{(r-1)}}\text{vec}(\mathbf{I}_N)\right\|^2+\lambda_T\mathbf{v}^{\text{T}}|\eta|,$$

where $\mathbf{v}=(|\widetilde{\xi}_1|^{-1},\ldots,|\widetilde{\xi}_{N^2}|^{-1})^{\text{T}}$, and remaining definitions are as in step 3. The above is also subjected to the constraints as stated in (3.6).

For the choice of λ_T and λ_T' used in the penalization of $\boldsymbol{\xi}$ and $\boldsymbol{\delta}$, respectively, we minimize the following BIC criterion with respect to both of them:

$$\begin{aligned}\text{BIC}(\lambda_{\xi,T},\lambda_{\delta,T}) &= \log\left(T^{-2}\left\|\mathbf{B}^{\text{T}}\mathbf{y}-\mathbf{B}^{\text{T}}\mathbf{Z}\widehat{\boldsymbol{\xi}}-\mathbf{B}^{\text{T}}\mathbf{Z}\mathbf{V}_0\widehat{\boldsymbol{\delta}}-\mathbf{B}^{\text{T}}\mathbf{X}_{\widehat{\boldsymbol{\beta}}}\text{vec}(\mathbf{I}_N)\right\|^2\right)\\ &\quad+|\widehat{S}|\frac{\log(T)}{T}\log(\log(2N-2))\end{aligned}\tag{4.17}$$

where $\widehat{\boldsymbol{\xi}}$ and $\widehat{\boldsymbol{\delta}}$ are the adaptive LASSO solutions with tuning parameters $\lambda_{\xi,T}$ and $\lambda_{\delta,T}$, respectively, and $\widehat{\boldsymbol{\beta}}=\boldsymbol{\beta}(\widehat{\boldsymbol{\xi}},\widehat{\boldsymbol{\delta}})$. The set $\widehat{S}$ is the indices for the non-zero values of $\widehat{\boldsymbol{\xi}}$. The BIC criterion (4.17) is in fact inspired by Wang et al. (2009).[4] Although the value of a is unknown in the definition of $\mathbf{B}$, the optimal values λ_T and λ_T' are in fact independent of a because of the logarithmic operation in the first term of (4.17). In practice we set $a=1$.

Remark 2. We mention in the discussions of Theorem 3 and 4 how to estimate $\boldsymbol{\Sigma}$ and $\mathbf{S}_0$ by finding a cut-off in the infinite sum. In practice, since the matrices involved in the infinite sum are usually almost zero everywhere when $|\tau|$ is larger than 2 or 3, we can compare the individual matrices at different τ, up to $|\tau|=10$ for example, and select a cut-off such that the sum changes little beyond that.

[4]We also experimented with the Extended BIC of Chen and Chen (2008) and obtained similar results.

4.1 Finding a more suitable $\boldsymbol{\gamma}$ and selection of instruments

We introduce $\boldsymbol{\gamma} = L^{-1}\mathbf{1}_L$ in Section 2.1 for aggregating the instruments. In fact, this can be estimated to provide maximal correlation with the endogenous variable $\mathbf{y}_t$ through two-stage least squares. In doing so, we consider the model

$$\mathbf{y}_t = \boldsymbol{\alpha} + \mathbf{B}_t\boldsymbol{\gamma} + \mathbf{v}_t,$$

where $\boldsymbol{\alpha}$ is an $N \times 1$ vector of unknown coefficients, and $\boldsymbol{\gamma}$ is the $K \times 1$ vector of coefficients we want to estimate. To get $\widehat{\boldsymbol{\gamma}}$, we can consider the problem

$$\min_{\boldsymbol{\alpha},\boldsymbol{\gamma}} \sum_{t=1}^{T} \left\|\mathbf{y}_t - \boldsymbol{\alpha} - \mathbf{B}_t\boldsymbol{\gamma}\right\|^2,$$

where we have the solution

$$\widehat{\boldsymbol{\gamma}} = \Big(\sum_{t=1}^{T}(\mathbf{B}_t - \bar{\mathbf{B}})^{\mathrm{T}}(\mathbf{B}_t - \bar{\mathbf{B}})\Big)^{-1} \sum_{t=1}^{T}(\mathbf{B}_t - \bar{\mathbf{B}})^{\mathrm{T}}(\mathbf{y}_t - \bar{\mathbf{y}}).$$

Implementing this does not change our proof much, since we can easily show that $\left\|\widehat{\boldsymbol{\gamma}}\right\|_1 = O_p(1)$, which substitutes $\left\|\boldsymbol{\gamma}\right\|_1 = 1$ in all of our proofs.

If the set (2.5) contains too many instruments, then we can follow Belloni et al. (2012) and use LASSO to select the most important linear combination of them, which essentially is a penalized version of the problem above. The proofs of all theorems still stay the same. Since in this paper we do not have many instruments, the practical performance in our simulations and real data analysis are indistinguishable among using $\boldsymbol{\gamma} = L^{-1}\mathbf{1}_L$, the least squares estimator $\widehat{\boldsymbol{\gamma}}$ above, or the LASSO estimator from Belloni et al. (2012).

5 Simulations and empirical illustration

We conduct a detailed Monte-Carlo exercise in Section 5.1 to demonstrate the finite-sample performance of our estimators. In Section 5.2, we use our methodology to analyze the determinants of cross-section variation in returns of the largest stocks traded on the New York Stock Exchange in 2017. We find that firms' stock returns exhibit a dependence in the cross-section that cannot be explained by observable characteristics such as similar sector or subsector of activity.

5.1 Simulation

We first construct the expert neighboring matrices from various measures of true distances between units. This approximates matrices that practitioners face in real-world applications. More specifically, we construct the expert matrices $\mathbf{W}_{0i}$ from the following ten measures of distance:

dummies for sharing a border; free-trade agreement; common currency; past colonial relation; common language, religion and origins of the legal system; and if the pair were ever part of the same nation. We further add non-binary relations for the inverse distance between capital cities and inverse time zone difference.[5] Finally, we construct the sparse deviation matrix $\mathbf{A}^*$ sampling 5% of its elements to be equal to 0.5. If any row sums of the ten $\mathbf{W}_{0i}$ or $\mathbf{A}^*$ exceed one, we divide by the L_1 norm of the row.

We construct the following scenarios to demonstrate the flexibility of the method. They exemplify the following situations to which the practitioner could be subjected to, and are particular cases of the general specification in Equation (2.3):

$$\mathbf{W}^* = \mathbf{A}^* + \sum_{i=1}^{M} \delta_i^* \cdot \mathbf{W}_{0i}. \tag{5.1}$$

(a) *No expert knowledge:* The practitioner either has no prior knowledge about the neighboring matrix or is unwilling to use it. This corresponds to the estimation of sparse spatial weight matrix without any prior specifications, and the method reduces to estimating the entire spatial weight matrix. In this case, we assume that $\mathbf{W}^* = \mathbf{A}^*$. The weights δ_i^* are not estimated, and $M = 0$.

(b) *Partial expert knowledge:* The practitioner uses the full information of the ten expert matrices, so $M = 10$, and selects which matrices best explain the data. Yet, the practitioner has partial information in the sense that he or she did not fully specify the expert matrices, and the true network has a sparse deviation from its linear combination. We assume that $\mathbf{W}^* = 0.2 \cdot \mathbf{W}_{01} + 0.2 \cdot \mathbf{W}_{02} + \mathbf{A}^*$, so $\delta_1^* = \delta_2^* = 0.2$ and $\delta_3^* = \cdots = \delta_{10}^* = 0$. The estimated model includes the $M = 10$ matrices.

(c) *Full expert knowledge:* Similar to the case above, except that we do not include the sparse deviation in the true network, i.e., $\mathbf{W}^* = 0.2 \cdot \mathbf{W}_{01} + 0.2 \cdot \mathbf{W}_{02}$, so $\delta_1^* = \delta_2^* = 0.2$ and $\delta_3^* = \cdots = \delta_{10}^* = 0$. The practitioner is uncertain if he or she has the correct specification, however, and includes the sparse deviation term in the estimated model.

By focusing on the case of no-expert knowledge, we can assess the performance in cases when no information is available about the true neighboring matrix. The partial and full information cases highlight the benefits of using the method combined with prior information on the neighboring matrices. Doing so, we make use of available information and estimate a model robust to misspecification or incompleteness.

At each period, disturbances $\boldsymbol{\epsilon}_t$ are jointly normally distributed, with variance-covariance matrix having 1 on the diagonal. Off-diagonal elements (i, j) above the main diagonal are randomly chosen to be 0.25 with 10% probability, and replicated to the (j, i) element below the main di-

[5] Downloaded from the GeoDist database from CEPII, available at `www.cepii.fr`.

agonal. This ensures that the variance-covariance matrix is symmetric and introduces spatial dependence also in the disturbance term. Each column of the covariates $\mathbf{X}_t$ are also generated as standard normals, and added to the $\boldsymbol{\epsilon}_t$ divided by half. Instruments $\mathbf{B}_t$ are equal to the covariates plus a standard normal noise. We provide simulations with alternative specifications in Subsection 1.1 of the supplementary materials to this article. The vector of true $\boldsymbol{\beta}^*$ is equal to one. All simulations introduce individual fixed effects. Finally, response data $\mathbf{y}_t$ are generated according to model (2.4). In all simulations, we consider $N = 25$, 50 and 75 and $T = 50$, 100 and 200. Tuning parameters are chosen by the BIC criterion in (4.17). Each combination has at least one thousand replications. Standard errors across replications are shown in parentheses.

In Table 1, we use several criteria to evaluate the performance of the estimator. Specificity is defined as the proportion of true zeros estimated as zeros. Sensitivity is the proportion of non-zeros estimated as non-zeros. These measures apply to $\mathbf{A}^*$ and $\boldsymbol{\delta}^*$. Furthermore, we present the L_1 norm $\left\|\widetilde{\boldsymbol{\xi}} - \boldsymbol{\xi}^*\right\|_1$ of the LASSO estimator $\widetilde{\boldsymbol{\xi}}$, and also $\left\|\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}^*\right\|_1$ of the adaptive LASSO estimator $\widehat{\boldsymbol{\xi}}$. We show the bias for the elements of $\mathbf{A}^*$, $\boldsymbol{\delta}^*$, and $\boldsymbol{\beta}^*$. Finally, we compute the sparsity of $\mathbf{A}^*$.

Table 1 shows very good performance of the estimator for various choices of N and T. Specificity of $\mathbf{A}^*$ is above 95% in most cases, and sensitivity approaches 100% as more time periods are made available. In the "no expert knowledge" case, performance is slightly better at lower T, and slightly deteriorates when expert matrices are included in the estimation in the "partial expert knowledge case." This is expected as sparse deviation may pick up estimation errors of $\boldsymbol{\delta}_0$. In the "full expert knowledge case," the estimated sparse deviation is 99% of zeros (since by construction there are no true non-zeros, the sensitivity is not defined). Compared to the earlier partial-knowledge case, the practitioner could then interpret this result as a correct specification of expert matrices.

Results on LASSO and adaptive LASSO L_1 norms show that the latter provided significant gains in performance. Moreover, the biases on $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ are small. Finally, selection of relevant $\boldsymbol{\delta}$'s is demonstrated by the specificity and sensitivity parameters, which in most cases are above 80%. Additional simulation results, which can be found in the supplementary materials to this paper, demonstrate the robustness of the procedure under alternative scenarios.

Table 1: Simulation results

	No knowledge			Partial knowledge			Full knowledge		
	T=50	T=100	T=200	T=50	T=100	T=200	T=50	T=100	T=200
.				**N = 25**					
$\mathbf{A}^*$ Specificity	.995	.998	.999	.995	.991	.977	.995	.996	.997
	(.004)	(.002)	(.001)	(.005)	(.008)	(.017)	(.005)	(.004)	(.005)
$\mathbf{A}^*$ Sensitivity	.975	1.000	1.000	.846	.933	.984	-	-	-
	(.030)	(.004)	(.000)	(.059)	(.053)	(.027)			
$\mathbf{A}^*$ bias	-.022	-.021	-.021	-.027	-.026	-.026	.000	.000	.000
	(.001)	(.000)	(.000)	(.002)	(.001)	(.002)	(.000)	(.000)	(.000)
LASSO L1	.036	.033	.030	.035	.035	.037	.013	.013	.010
	(.002)	(.002)	(.001)	(.003)	(.003)	(.004)	(.002)	(.002)	(.001)
AdaLASSO L1	.023	.022	.022	.028	.027	.027	.000	.000	.000
	(.001)	(.000)	(.000)	(.002)	(.001)	(.002)	(.000)	(.000)	(.000)
Sparsity	.946	.948	.949	.953	.945	.929	.995	.996	.997
	(.005)	(.002)	(.001)	(.005)	(.008)	(.016)	(.005)	(.004)	(.005)
$\boldsymbol{\beta}$ bias	.019	.015	.011	.038	.033	.029	.029	.020	.014
	(.011)	(.008)	(.006)	(.020)	(.016)	(.013)	(.016)	(.010)	(.007)
$\boldsymbol{\delta}^*$ Specificity	1.000	1.000	1.000	.876	.814	.751	.999	1.000	1.000
	(.000)	(.000)	(.000)	(.085)	(.095)	(.088)	(.009)	(.000)	(.000)
$\boldsymbol{\delta}^*$ Sensitivity	-	-	-	.750	.816	.944	.783	.839	.905
				(.270)	(.246)	(.158)	(.162)	(.167)	(.151)
$\boldsymbol{\delta}^*$ Bias	.000	.000	.000	.006	.015	.027	-.021	-.016	-.011
	(.000)	(.000)	(.000)	(.013)	(.014)	(.011)	(.007)	(.007)	(.007)
				N = 50					
$\mathbf{A}^*$ Specificity	.960	.972	.984	.958	.953	.940	.961	.970	.977
	(.004)	(.004)	(.003)	(.010)	(.014)	(.018)	(.004)	(.005)	(.007)
$\mathbf{A}^*$ Sensitivity	.872	.980	1.000	.666	.847	.966	-	-	-
	(.031)	(.013)	(.002)	(.083)	(.058)	(.020)			
$\mathbf{A}^*$ bias	-.012	-.011	-.011	-.018	-.017	-.016	.000	.000	.000
	(.000)	(.000)	(.000)	(.002)	(.002)	(.001)	(.000)	(.000)	(.001)
LASSO L1	.029	.027	.023	.029	.028	.028	.017	.016	.013
	(.001)	(.001)	(.001)	(.001)	(.001)	(.001)	(.001)	(.001)	(.001)
AdaLASSO L1	.015	.013	.011	.019	.018	.017	.002	.002	.001
	(.001)	(.000)	(.000)	(.002)	(.002)	(.001)	(.000)	(.000)	(.000)
Sparsity	.918	.925	.934	.927	.913	.894	.961	.970	.977
	(.004)	(.003)	(.003)	(.007)	(.012)	(.016)	(.004)	(.005)	(.007)
$\boldsymbol{\beta}$ bias	.010	.009	.007	.034	.026	.020	.025	.017	.011
	(.006)	(.005)	(.004)	(.018)	(.012)	(.009)	(.012)	(.009)	(.006)
$\boldsymbol{\delta}^*$ Specificity	1.000	1.000	1.000	.770	.749	.714	.991	.988	.989
	(.000)	(.000)	(.000)	(.091)	(.095)	(.090)	(.035)	(.039)	(.039)
$\boldsymbol{\delta}^*$ Sensitivity	-	-	-	.994	.998	.998	.895	.898	.917
				(.053)	(.030)	(.030)	(.155)	(.154)	(.144)
$\boldsymbol{\delta}^*$ Bias	.000	.000	.000	.040	.039	.040	-.009	-.004	.001
	(.000)	(.000)	(.000)	(.011)	(.010)	(.006)	(.010)	(.013)	(.015)
				N = 75					
$\mathbf{A}^*$ Specificity	.945	.958	.967	.931	.933	.995	.998	.999	.999
	(.003)	(.003)	(.002)	(.007)	(.010)	(.004)	(.001)	(.001)	(.001)
$\mathbf{A}^*$ Sensitivity	.747	.879	.989	.750	.790	.805	-	-	-
	(.032)	(.021)	(.007)	(.060)	(.048)	(.096)			
$\mathbf{A}^*$ bias	-.009	-.008	-.008	-.012	-.012	-.010	.000	.000	.000
	(.000)	(.000)	(.000)	(.002)	(.001)	(.001)	(.000)	(.000)	(.000)
LASSO L1	.024	.024	.023	.025	.025	.017	.008	.008	.009
	(.000)	(.000)	(.000)	(.001)	(.001)	(.001)	(.000)	(.001)	(.001)
AdaLASSO L1	.013	.010	.009	.015	.013	.011	.000	.000	.000
	(.000)	(.000)	(.000)	(.001)	(.001)	(.001)	(.000)	(.000)	(.000)
Sparsity	.916	.916	.920	.907	.897	.955	.998	.999	.999
	(.003)	(.003)	(.002)	(.005)	(.008)	(.006)	(.001)	(.001)	(.001)
$\boldsymbol{\beta}$ bias	.006	.006	.005	.031	.021	.013	.017	.011	.008
	(.003)	(.004)	(.003)	(.016)	(.011)	(.006)	(.009)	(.006)	(.004)
$\boldsymbol{\delta}^*$ Specificity	1.000	1.000	1.000	.729	.709	.790	1.000	1.000	1.000
	(.000)	(.000)	(.000)	(.107)	(.097)	(.092)	(.000)	(.000)	(.000)
$\boldsymbol{\delta}^*$ Sensitivity	-	-	-	1.000	1.000	1.000	.729	.828	.865
				(.000)	(.000)	(.000)	(.131)	(.167)	(.164)
$\boldsymbol{\delta}^*$ Bias	.000	.000	.000	.040	.043	.027	-.022	-.017	-.013
	(.000)	(.000)	(.000)	(.013)	(.010)	(.011)	(.004)	(.006)	(.007)

Notes: Simulated results under various combinations of N and T for 1,000 iterations. "No knowledge case" refers to the "No expert knowledge case,"where expert matrices are not used in the estimated model, and the true network is defined by the sparse deviation only. In the "Partial knowledge" case, the true matrix is a combination of two expert matrices and a sparse deviation. There are no sparse deviations in true matrix of the "Full knowledge" case, but it is included in the estimated model. Specificity (Sensitivity) refers to the proportion of true zeros (non-zeros) that are estimated as zeros (non-zeros). Lasso L1 and AdaLasso L1 refer to the L_1 norm of the vectorized sparse deviation matrix of the LASSO and adaptive LASSO steps, respectively. Standard error across are calculated across iterations. Penalization parameters are chosen by BIC.

5.2 Empirical illustration

In this subsection, we illustrate how the procedure can be applied to uncover new findings in empirical practice. We provide suggestive evidence that the stock returns exhibit significant cross-sectional correlation, even after controlling for fluctuation in the market risk to account for the co-movement in stocks due to external reasons to specific pairs of stocks. Thus the residual correlation in stocks can neither be fully explained by aggregate shocks nor by several potential measures of proximity between firms, both economical and geographical.

Fama and French (1992, 1993 and 1996) proposed a benchmark stock return model which include factors to capture movements in market risk. The so-called "Fama-French factors" are calculated over a very large class of stocks. These factors are meant to absorb co-movements that are not particular to a specific pair of stocks, but are instead due to general market fluctuations. See Feng et al. (2017) for additional factors and measures of market risk. However, these papers did not explore that, beyond general market movements, stock returns might directly affect other stock returns. One possibility, for example, is that firms are subject to more specific sectoral or state shocks, or that interconnectedness in the supply chains is reflected as co-movement in stock returns. Other papers in the literature, such as Engle et al. (2012), Diebold and Yilmaz (2015), and recent work by Barigozzi and Brownlees (2018) consider the spillovers of the *volatility* measures across markets.[6]

To bring light to this issue and quantify the prevalence of cross-section dependence in the intra-market stock returns, we build a panel of daily returns of the largest $N = 75$ firms traded on the New York Stock Exchange throughout 2017. We obtain $T = 251$ trading days. Reproducing the Fama-French papers, we use their factors as covariates.[7] We consider the following eight measures of proximity: firms' same sector and subsector of activity according to GICS classifications, state and city of the headquarters, inverse and inverse squared distance, and the state-sector and subsector-state interactions. If any row sums of the matrices exceed one, we divide them by the L_1 norm of the row. These constitute our expert matrices $\mathbf{W}_{0i}$.

In specification *(i)* we make no use of expert matrices, corresponding to "no expert knowledge" in the previous subsection. The density of the estimated sparse adjustment matrix (defined as one minus the sparsity) is 12.4%. It is not significantly affected by the inclusion of three or five Fama-French factors. We then add the two measures of economic distance in specification *(ii)*: similarity along sectors or subsectors of activity. Subsector of activity does explain the dependence structure in the cross-section of returns. These results are robust to the inclusion of additional Fama-French factors. This is intuitive, as firms in the same subsector of activity can be subject to common shocks. We then explore measures of distance related to geography in specification *(iii)*. Geographical distance by itself does not matter in any specification. Finally,

[6] For other work in this are, see also Kutzker and Wied (2018).

[7] Daily data for the Fama-French factors are available at French's data library at `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`.

in *(iv)* we introduce the interaction of geography and sectors and subsector of activity. In one specification, the interaction of state and sector of activity has explanatory power. Again, this is intuitive, as firms in the same sector of activity may be subject to common state-specific shocks. All specifications contain stock fixed effects.

The density of the estimated sparse adjustment matrix fell from 12.4% to 6.7% in the less parsimonious specification *(iv)*. This indicates that the cross-sectional correlation in stocks may be partly explained by observable characteristics of the firms, but not to the full extent to lead to an adjustment matrix $\mathbf{A}^*$ populated only by zeros.

Finally, by interpreting the estimated sparse adjustment matrix, we observe a similar reduction in the clustering coefficient. This coefficient measures the extent to which stock returns form small co-moving subgroups. More specifically, it is defined as the fraction of connected triads that are triangles; that is, the number of times where the (i, j)-th, (j, k)-th and (k, i)-th elements are non-zero under the estimated $\mathbf{A}^*$ over the number of times that only the (i, j)-th and (j, k)-th elements are non-zero. Therefore, a smaller (larger) coefficient can be interpreted as less (more) co-movement along small groups of stocks. We find that the clustering coefficients fall by about half, from .026 in specification *(i)* to .011 in specification *(iv)*. This is in line with the expectation that groups were, to a large extent, defined by the combination of economic and geographic distance.

Overall, these results point out the importance of idiosyncratic elements in the co-movement of stocks, captured in the sparse adjustment matrix, that cannot be explained by Fama-French factors or economical or geographical measures of proximity between firms.

Table 2: Dependence in stock returns

	3 Fama-French factors				**5 Fama-French factors**			
	(i)	(ii)	(iii)	(iv)	(i)	(ii)	(iii)	(iv)
Excess return (Rm-Rf)	.011	.006	.003	.003	.012	.006	.005	.005
	(.001)	(.001)	(.005)	(.005)	(.001)	(.001)	(.004)	(.004)
Small minus Big (SMB)	-.009	-.004	-.002	-.002	-.008	-.004	-.003	-.003
	(.002)	(.001)	(.007)	(.007)	(.002)	(.001)	(.006)	(.006)
High minus Low (HML)	.013	.005	.001	.001	.012	.004	.001	.001
	(.002)	(.001)	(.007)	(.007)	(.002)	(.001)	(.006)	(.006)
Robust minus Weak (RMW)					.005	.001	.001	.001
					(.002)	(.001)	(.007)	(.007)
Conservative minus Aggressive (CMA)					.003	.002	.002	.002
					(.002)	(.001)	(.006)	(.006)
Sector	-	.000	.000	.000	-	.000	.000	.000
Subsector		.161	.173	.245		.120	.112	.055
		(.025)	(.026)	(.036)		(.031)	(.034)	(.023)
State			.000	.000			.000	.000
City			.000	.000			.000	.000
Distance^(-1)			.000	.000			.000	.000
Distance^(-2)			.000	.000			.000	.000
Sector x state				.000				.107
								(.023)
Subsector x state				.000				.000
Density	.124	.122	.077	.076	.124	.122	.068	.067
Clustering	.026	.025	.013	.012	.025	.025	.012	.011
Size of largest component	75	75	75	75	75	75	75	75

Notes: Estimated model results with three and five Fama-French factors (Fama and French, 1993) for the cross-section of stock returns of the $N = 75$ stocks with largest trading volume in 2017. $T = 251$. The five Fama-French factors are: the excess return on the market (Rm-Rf) and measures of dispersion as captured by the average return on a portfolio of small stock minus big stocks (Small minus Big, SMB), the average return on the value portfolios minus the average return on growth portfolios (High minus Low, HML), the average return of operating profitability portfolios minus the return on weakly-operating return portfolios (Robust minus Weak, RMW) and, finally, the difference between the return in conservative and aggressive investment portfolios (Conservative minus Agressive, CMA). The first panel shows the estimated $\boldsymbol{\beta}$ coefficient with standard errors in parenthesis. The second panel shows the estimated $\boldsymbol{\delta}$ coefficients for four specifications: with no expert matrix, with economic expert matrices (sector and subsector of activity), with geographic proximity matrices (state, city, inverse and inverse squared distance, calculated over the address of the headquarters), and the interaction of economic and state. Density is one minus sparsity. Clustering coefficient is the fraction of connected triads that are triangles. The size of the largest coefficient is the number of elements of the smallest submatrix such that every stock is connected at least one other, possibly by paths of any length. In our case, the size of the largest component is always equal to the number of stocks.

6 Conclusion

In this paper, we unify the selection and estimation of a spatial weight matrix in a spatial autoregressive model through the introduction of a sparse adjustment matrix, added to a linear combination of specified spatial weight matrices from expert knowledge. Without any expert knowledge, the problem reduces to pure spatial weight matrix estimation. When one or more spatial matrices are used, this is a selection plus estimation problem. The estimation of the sparse adjustment matrix, and the selection of which specified spatial weight matrix to include in a linear combination, are done through solving two respective adaptive LASSO problems. Theoretical results support inferences on various parameters including the elements in the sparse adjustment matrix itself, with practical implementation also discussed.

From the simulations and real data analysis, we see that our method indeed practically allows for the improvement of the spatial weight matrix estimation through giving a non-zero estimated sparse adjustment matrix in the stock returns example. This provides insights into the co-movements of different spatial units and how much our expert knowledge, translated into specified spatial weight matrices, helps in understanding such co-movements.

Supplementary Material

All the proofs are presented in the supplementary materials in this paper.

References

Ahrens, A. and A. Bhattacharjee (2015). Two-step lasso estimation of the spatial weights matrix. *Econometrics 3*(1), 128.

Ammermuller, A. and J.-S. Pischke (2009). Peer effects in european primary schools: Evidence from pirls. *Journal of Labor Economics 27*(3), 315–348.

Angrist, J. D. and K. Lang (2004, December). Does school integration generate peer effects? evidence from boston's metco program. *American Economic Review 94*(5), 1613–1634.

Arnold, M., S. Stahlberg, and D. Wied (2011). Modeling different kinds of spatial dependence in stock returns. *Empirical Economics 44*(2), 761–774.

Bai, Z. and J. Silverstein (2009). *Spectral Analysis of Large Dimensional Random Matrices*. Springer Series in Statistics. Springer New York.

Bailey, N., S. Holly, and M. H. Pesaran (2016). A two stage approach to spatio temporal analysis with strong and weak cross sectional dependence. *Journal of Applied Econometrics 31*(1), 249–280.

Bańbura, M., D. Giannone, and L. Reichlin (2010). Large bayesian vector auto regressions. *Journal of Applied Econometrics 25*(1), 71–92.

Barigozzi, M. and C. T. Brownlees (2018). Nets: Network estimation for time series.

Beenstock, M. and D. Felsenstein (2012). Nonparametric estimation of the spatial connectivity matrix using spatial panel data. *Geographical Analysis 44*(4), 386–397.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica 80*(6), 2369–2429.

Bernanke, B. S., J. Boivin, and P. Eliasz (2005). Measuring the effects of monetary policy: a factor-augmented vector autoregressive (favar) approach. *The Quarterly journal of economics 120*(1), 387–422.

Bhattacharjee, A. and C. Jensen-Butler (2013). Estimation of the spatial weights matrix under structural constraints. *Regional Science and Urban Economics 43*(4), 617 – 634.

Chen, J. and Z. Chen (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika 95*(3), 759–771.

Chen, X., M. Xu, and W. B. Wu (2013, 12). Covariance and precision matrix estimation for high-dimensional time series. *The Annals of Statistics 41*(6), 2994–3021.

Corrado, L. and B. Fingleton (2011). Multilevel modelling with spatial effects.

Dicker, L., B. Huang, and X. Lin (2013). Variable selection and estimation with the seamless-l0 penalty. *Statistica Sinica 23*, 929–962.

Diebold, F. X. and K. Yilmaz (2015). Trans-atlantic equity volatility connectedness: Us and european financial institutions, 2004–2014. *Journal of Financial Econometrics 14*(1), 81–127.

Elhorst, J. P. (2014). Dynamic spatial panels: models, methods and inferences. In *Spatial Econometrics*, pp. 95–119. Springer.

Engle, R. F., G. M. Gallo, and M. Velucchi (2012). Volatility spillovers in east asian financial markets: a mem-based approach. *Review of Economics and Statistics 94*(1), 222–223.

Fama, E. F. and K. R. French (1992). The cross-section of expected stock returns. *the Journal of Finance 47*(2), 427–465.

Fama, E. F. and K. R. French (1993). Common risk factors in the returns on stocks and bonds. *Journal of financial economics 33*(1), 3–56.

Fama, E. F. and K. R. French (1996). Multifactor explanations of asset pricing anomalies. *The journal of finance 51*(1), 55–84.

Fan, J. and Y. Liao (2014, 06). Endogeneity in high dimensions. *Ann. Statist. 42*(3), 872–917.

Fan, J. and J. Lv (2011). Non-concave penalized likelihood with np-dimensionality. *IEEE Transactions on Information Theory 57*, 5467–5484.

Feng, G., S. Giglio, and D. Xiu (2017). Taming the factor zoo.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*, 1–22.

Glaeser, E., B. Sacerdote, and J. A. Scheinkman (1996, May). Crime and social interactions. *The Quarterly Journal of Economics 111*(2), 507–548.

Kelejian, H. H. and G. Piras (2011). An extension of kelejian's j test for non-nested spatial models. *Regional Science and Urban Economics 41*(3), 281 – 292.

Kelejian, H. H. and G. Piras (2016). An extension of the j test to a spatial panel data framework. *Journal of Applied Econometrics 31*(2), 387–402.

Kelejian, H. H. and I. R. Prucha (1998). A generalized spatial two stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *Journal of Real Estate Finance and Economics 17*, 99–121.

Kock, A. B. and L. Callot (2015). Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics 186*(2), 325–344.

Kutzker, T. and D. Wied (2018). Testing the correct specification of a spatial dependence panel model for stock returns.

Lam, C. and P. C. L. Souza (2015). Detection and estimation of block structure in spatial weight matrix. *Econometric Reviews*, 1–30.

Lee, L.-F. and J. Yu (2010). Estimation of spatial autoregressive panel data models with fixed effects. *Journal of Econometrics 154*, 165–185.

LeSage, J. and R. K. Pace (2008). *Introduction to Spatial Econometrics*. Chapman and Hall.

Liu, X. and I. R. Prucha (2017, August). A robust test for network generated dependence. Working paper.

Longstaff, F. A. (2010, September). The subprime credit crisis and contagion in financial markets. *Journal of Financial Economics 97*(3), 436–450.

Medeiros, M. and E. Mendes (2016). l1-regularization of high-dimensional time-series models with non-gaussian and heteroskedastic errors. *Journal of Econometrics 191*(1), 255–271.

Pesaran, M. H., T. Schuermann, and S. M. Weiner (2004). Modeling regional interdependencies using a global error-correcting macroeconometric model. *Journal of Business & Economic Statistics 22*(2), 129–162.

Pinkse, J., M. E. Slade, and C. Brett (2002). Spatial price competition: A semiparametric appraoch. *Econometrica 70*(3), 1111–1153.

Shao, X. (2010). Nonstationary-extended whittle estimation. *Econometric Theory 26*, 1060–1087.

Sun, Y. (2016). Functional-coefficient spatial autoregressive models with nonparametric spatial weights. *Journal of Econometrics 195*(1), 134–153.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58*(1), 267–288.

Wang, H., B. Li, and C. Leng (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(3), 671–683.

Wu, W. B. (2005). Nonlinear system theory: another look at dependence. *Proc. Natl. Acad. Sci. USA 102*, 14150–14154.

Wu, W. B. (2011). Asymptotic theory for stationary processes. *Statistics and Its Interface 4*, 207–226.

Zhao, P. and B. Yu (2006). On model selection consistency of lasso. *Journal of Machine Learning Research 7*, 2541–2563.

Zhou, Z. (2010). Nonparametric inference of quantile curves for nonstationary time series. *Ann. Statist. 38*(4), 2187–2217.

Zou, H. (2006, December). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*, 1418–1429.